



University  
of Glasgow

Carr, David William (2016) *Development of statistical methods for composite environmental quality indices at data zone resolution*. MSc(R) thesis.

<http://theses.gla.ac.uk/7280/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given



University  
of Glasgow

# Development of Statistical Methods for Composite Environmental Quality Indices at Data Zone Resolution

David William Carr

*Submitted in fulfilment of the requirements for the Degree of Master of Science*

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

April 2016

# Abstract

The principal objective of this thesis is to develop methods for composite indices measuring environmental quality at small spatial scales. Specifically, it is of interest to create a composite environmental quality index at data zone resolution in Greater Glasgow, where data zones are small geographical areas defined by the Scottish Government and used to report various statistics and indicators. The index will consist of various indicators measuring different aspects of environmental quality, being grouped into three ‘domains’: air quality, soil quality and water quality.

Composite indices are multidimensional summaries of data, constructed from a number of one-dimensional indicators measuring one variable each. This is primarily achieved through converting the scales of these individual indicators to a common, unit-less scale and then aggregating the indicators together to form a composite index. Each indicator can be allocated an equal weighting within the composite index or weightings can be allowed to vary. Various subjective choices throughout the construction process can affect a composite index, all of which must be considered when interpreting the final result.

Chapter 1 provides background information on environmental issues such as air, soil and water quality in urban areas and the use of composite indices in several contexts, particularly in environmental situations. A literature review of composite index construction methodology is also presented with much of the methodology informing the work of later chapters.

Chapter 2 introduces the various data sets that will be used to construct a composite environmental quality index for Greater Glasgow. Exploratory spatial and temporal analyses of these data will also be detailed prior to any formal statistical modelling.

Chapter 3 discusses the fitting of geostatistical models to the data. These models will then be used to predict the various environmental processes at a high enough spatial resolution for there to be estimated values for each data zone in the study region.

Chapter 4 details how the modelled data from Chapter 3 will be used to construct separate air, soil and water quality indices and then aggregating these domains to create a general, multi-dimensional environmental quality index. A variety of different methods found in the literature will be used and the indices will be reported at data zone resolution.

Chapter 5 will assess the composite indices constructed in Chapter 4 by investigating possible methods for determining the robustness and usefulness of the indices. Firstly, an uncertainty analysis of the various indices will be considered to quantify how much statistical variability would be expected for each data zone. Various approaches for assessing how the index can effectively capture changes in environmental quality over time will then be investigated. This will be restricted to a period of five years.

Chapter 6 consists of a final discussion of the results presented in this thesis as well as providing suggestions for any further work in this area.

# Contents

<b>Abstract</b>	<b>i</b>
<b>List of Tables</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>Acknowledgements</b>	<b>xiv</b>
<b>1 Background</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.1.1 Examples of Past and Current Environmental Indices . . . . .	2
1.2 Aspects of Environmental Quality . . . . .	4
1.2.1 Air . . . . .	4
1.2.2 Water . . . . .	7
1.2.3 Land . . . . .	8
1.3 Construction of Composite Indices . . . . .	9
1.4 Overview and Aims of Thesis . . . . .	27
<b>2 Data and Exploratory Analysis</b>	<b>29</b>
2.1 Air . . . . .	29
2.1.1 Air Pollutants . . . . .	29
2.1.2 Air Quality Monitoring Station Data . . . . .	30
2.1.3 DEFRA Modelled Grids . . . . .	34
2.1.4 $NO_2$ Diffusion Tube Data . . . . .	39

2.1.5	Exploratory Spatial Analysis of All Air Quality Data . . . . .	41
2.2	Water . . . . .	43
2.2.1	Water Determinands . . . . .	43
2.2.2	Exploratory Analysis . . . . .	47
2.3	Land . . . . .	51
2.3.1	Soil Pollutants . . . . .	52
2.3.2	Exploratory Analysis . . . . .	54
2.4	Temporal Frequency of Data . . . . .	55
2.5	Chapter Summary . . . . .	56
<b>3</b>	<b>Geostatistical Modelling of Air, Water and Soil Data</b>	<b>59</b>
3.1	Methods . . . . .	59
3.1.1	Flexible Regression . . . . .	59
3.1.2	Geostatistical Modelling . . . . .	66
3.1.3	Model Comparison . . . . .	70
3.1.4	Random Forests . . . . .	71
3.2	Modelling Continuous Environmental Determinands . . . . .	73
3.2.1	Soil Quality Models . . . . .	73
3.2.2	Air Quality Models . . . . .	76
3.2.3	Water Quality Models for Chemical Determinands . . . . .	79
3.3	Macroinvertebrate Analysis Using Random Forests . . . . .	83
3.4	Prediction and Aggregation to Data Zone Level . . . . .	84
3.5	Chapter Summary . . . . .	86
<b>4</b>	<b>Development of Composite Indices</b>	<b>87</b>
4.1	Methods . . . . .	87
4.1.1	Index Construction Steps . . . . .	87
4.1.2	Factor Analysis . . . . .	92
4.2	Domain Indices . . . . .	94
4.2.1	Air Quality Index . . . . .	94

4.2.2	Soil Quality Index . . . . .	96
4.2.3	Water Quality Index . . . . .	100
4.3	Environmental Quality Index . . . . .	103
4.4	Tabular Summary of Results . . . . .	106
4.5	Chapter Summary . . . . .	106
<b>5</b>	<b>Assessment of Composite Indices</b>	<b>110</b>
5.1	Introduction . . . . .	110
5.2	Methods . . . . .	112
5.2.1	Monte Carlo Sampling . . . . .	112
5.2.2	Kappa Statistics . . . . .	114
5.3	Results . . . . .	116
5.3.1	Uncertainty Analysis . . . . .	116
5.3.2	Comparing Indices Over Time . . . . .	120
5.4	Chapter Summary . . . . .	127
<b>6</b>	<b>Discussion and Conclusions</b>	<b>130</b>
6.1	Further Work . . . . .	134
	<b>Appendices</b>	<b>138</b>
A	Additional Material Related to Chapter 2 . . . . .	138
B	Additional Material Related to Chapter 3 . . . . .	160
	<b>Bibliography</b>	<b>171</b>

# List of Tables

1.1	Indicator framework for environmental protection (Eurostat, 2014) . . .	4
1.2	Steps for constructing a composite index (OECD, 2008) . . . . .	10
1.3	Structure of the Environmental Performance Index (Hsu et al., 2014)	21
1.4	Selection criteria for data in the Environmental Performance Index (Hsu et al., 2014) . . . . .	21
2.1	Descriptive statistics for selected $NO_2$ air quality monitoring stations in 2011 (units of measurement = $\mu g/m^3$ ) . . . . .	32
2.2	Descriptive statistics for DEFRA modelled grids (2011 – units of mea- surement = $\mu g/m^3$ ) . . . . .	39
2.3	Descriptive statistics for $NO_2$ diffusion tube data (units of measure- ment = $\mu g/m^3$ ) . . . . .	41
2.4	Descriptive statistics for water chemistry determinands . . . . .	47
2.5	Status and catchment information for macroinvertebrate data . . . .	51
2.6	Soil determinands and limits of detection (British Geological Survey, 2014) – units of measurement = mg/kg . . . . .	52
2.7	Descriptive statistics of metals in G-BASE data set (units of measure- ment = mg/kg) . . . . .	54
2.8	Temporal frequency of data . . . . .	56
3.1	Results of flexible regression model for lead . . . . .	74
3.2	Results of flexible regression model for $NO_2$ . . . . .	77
3.3	Results of flexible regression model for phosphorus . . . . .	80
3.4	Results of analysis of variance for flexible regression model . . . . .	81



3.5	Classification matrix for random forest model . . . . .	83
3.6	Variable importance for random forest model . . . . .	83
4.1	Indicator weightings for air quality index (AQI) . . . . .	94
4.2	Indicator weightings for soil quality index (SQI) . . . . .	97
4.3	Indicator weightings for water quality index (WQI) . . . . .	100
4.4	Domain weightings for environmental quality index (EQI) . . . . .	103
4.5	Ranked index values for eleven selected data zones (1 = best environ- mental quality; 1748 = worst environmental quality) . . . . .	107
4.6	Z-score index values for eleven selected data zones . . . . .	107
4.7	Re-scaled index values for eleven selected data zones (0 = best envi- ronmental quality; 1 = worst environmental quality) . . . . .	108
5.1	Categories for interpreting Cohen's kappa statistic (Vierra and Gar- rett, 2005) . . . . .	116

# List of Figures



1.1	Indicators, aggregation methods and weightings used by the Scottish Index of Multiple Deprivation (Scottish Government, 2012) . . . . .	17
1.2	Map of study region with local authority and data zone boundaries – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	28
2.1	Maps of air monitoring stations – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	31
2.2	(a) histogram of $NO_2$ concentration ( $\mu\text{g}/\text{m}^3$ ) values; (b) histogram of log-transformed $NO_2$ concentration ( $\mu\text{g}/\text{m}^3$ ) values . . . . .	33
2.3	Time series plots of log-transformed $NO_2$ concentrations ( $\mu\text{g}/\text{m}^3$ ) at selected monitoring stations . . . . .	35
2.4	Boxplots of monthly log-transformed $NO_2$ concentrations ( $\mu\text{g}/\text{m}^3$ ) at selected monitoring stations . . . . .	36
2.5	Boxplots of day-within-week log-transformed $NO_2$ concentrations ( $\mu\text{g}/\text{m}^3$ ) at selected monitoring stations . . . . .	37
2.6	Boxplots of hourly log-transformed $NO_2$ concentrations ( $\mu\text{g}/\text{m}^3$ ) at selected monitoring stations . . . . .	38
2.7	Maps with average 2011 $NO_2$ concentrations ( $\mu\text{g}/\text{m}^3$ ) for each 1 km by 1 km grid . . . . .	39
2.8	Maps with average 2011 $PM_{10}$ concentrations ( $\mu\text{g}/\text{m}^3$ ) for each 1 km by 1 km grid . . . . .	40

2.9	Map of $NO_2$ diffusion tube locations in Greater Glasgow – contains Ordnance Survey data © Crown copyright and database right (2014)	41
2.10	Boxplots of log-transformed $NO_2$ concentrations ( $\mu\text{g}/\text{m}^3$ ) by year . . .	42
2.11	(a) scatterplot of easting versus $NO_2$ concentration ( $\mu\text{g}/\text{m}^3$ ); (b) scatterplot of easting versus log-transformed $NO_2$ concentration ( $\mu\text{g}/\text{m}^3$ ); (c) scatterplot of northing versus $NO_2$ concentration ( $\mu\text{g}/\text{m}^3$ ); (d) scatterplot of northing versus log-transformed $NO_2$ concentration ( $\mu\text{g}/\text{m}^3$ )	42
2.12	Map of data collection points – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	44
2.13	(a) plot of Easting versus phosphorus concentration (g/L); (b) plot of Easting versus log-transformed phosphorus concentration (g/L); (c) plot of phosphorus concentration (g/L) versus Northing; (d) plot of log-transformed phosphorus concentration (g/L) versus Northing . . .	48
2.14	(a) boxplots of annual log-transformed phosphorus concentration (g/L); (b) boxplots of monthly log-transformed phosphorus concentration (g/L) . . . . .	48
2.15	(a) boxplots of annual log-transformed phosphorus concentration (g/L); (b) boxplots of monthly log-transformed phosphorus concentration (g/L) . . . . .	49
2.16	Annual proportions of macroinvertebrate classifications (using the average score per taxon) . . . . .	50
2.17	Map of macroinvertebrate sampling locations and classifications (● = bad, ● = poor, ● = moderate, ● = good, ● = high) – contains Ordnance Survey data © Crown copyright and database right (2014)	50
2.18	Map of sites for G-BASE project (Fordyce et al., 2012) – contains Ordnance Survey data © Crown copyright and database right (2014)	52
2.19	(a) scatterplot of easting versus lead concentration (mg/kg); (b) scatterplot of easting versus log-transformed lead concentration (mg/kg); (c) scatterplot of northing versus lead concentration (mg/kg); (d) scatterplot of northing versus log-transformed lead concentration (mg/kg)	55

2.20	(a) histogram of lead concentration (mg/kg) values; (b) histogram of log-transformed lead concentration (mg/kg) values . . . . .	56
3.1	Theoretical semi-variogram (University of Edinburgh, 2015) . . . . .	69
3.2	(a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals . .	75
3.3	Variogram for lead flexible regression model residuals . . . . .	75
3.4	Prediction surface for flexible regression model for log-transformed lead concentration (mg/kg) with sampling locations . . . . .	76
3.5	(a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals . .	78
3.6	Variogram for $NO_2$ additive model residuals . . . . .	78
3.7	Prediction surface for flexible regression model for log-transformed $NO_2$ concentration ( $\mu g/m^3$ ) with sampling locations . . . . .	79
3.8	(a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals . .	81
3.9	Prediction surface for flexible regression model for phosphorus concentration with sampling locations . . . . .	82
3.10	Prediction surface for flexible regression model for log-transformed phosphorus concentration (g/L) with sampling locations . . . . .	82
4.1	Flow chart detailing general steps in constructing composite indices .	88
4.2	Air quality index (AQI) for 2011 using ranking method (1 = best air quality; 1748 = worst air quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	95
4.3	Air quality index (AQI) for 2011 using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014)	96

4.4	Air quality index (AQI) for 2011 using re-scaling method (0 = best air quality; 1 = worst air quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	97
4.5	Soil quality index (SQI) using ranking method (1 = best soil quality; 1748 = worst soil quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	98
4.6	Soil quality index (SQI) using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	99
4.7	Soil quality index (SQI) using re-scaling method (0 = best soul quality; 1 = worst soil quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	99
4.8	Water quality index (WQI) for 2011 using ranking method (1 = best water quality; 1748 = worst water quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	101
4.9	Water quality index (WQI) for 2011 using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014)	102
4.10	Water quality index (WQI) for 2011 using re-scaling method (0 = best water quality; 1 = poorest water quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	102
4.11	Environmental quality index (EQI) for 2011 using ranking method (1 = best environmental quality; 1748 = worst environmental quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	103
4.12	Environmental quality index (EQI) for 2011 using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	104
4.13	Environmental quality index (EQI) for 2011 using re-scaling method (0 = best environmental quality; 1 = worst environmental quality) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	105

4.14	Location of eleven selected data zones in Greater Glasgow (those included are coloured red) . . . . .	106
5.1	5 <sup>th</sup> and 95 <sup>th</sup> percentile intervals for ranking normalisation method: (a) factor analysis weights; (b) equal weights (Legend: ▲ = EQI rank; ○ = median) . . . . .	117
5.2	5 <sup>th</sup> and 95 <sup>th</sup> percentile intervals for re-scaling normalisation method: (a) factor analysis weights; (b) equal weights (Legend: ▲ = EQI rank; ○ = median) . . . . .	118
5.3	5 <sup>th</sup> and 95 <sup>th</sup> percentile intervals for z-score normalisation method: (a) factor analysis weights; (b) equal weights (Legend: ▲ = EQI rank; ○ = median) . . . . .	119
5.4	Time series plots (2008 – 2012) for DZ S01003029: (a) EQI for ranking method with uncertainty intervals; (b) Ranked EQI change over time (2008 = 0); (c) EQI for re-scaling method with uncertainty intervals; (d) Re-scaled EQI change over time (2008 = 0); (e) EQI for z-score method with uncertainty intervals; (e) EQI for re-scaling method with uncertainty intervals; (f) Z-score EQI change over time (2008 = 0) (Legend: — = 95 <sup>th</sup> percentile; — = 5 <sup>th</sup> percentile) . . . . .	122
5.5	Time series plots (2008 – 2012) for DZ S01004896: (a) EQI for ranking method with uncertainty intervals; (b) Ranked EQI change over time (2008 = 0); (c) EQI for re-scaling method with uncertainty intervals; (d) Re-scaled EQI change over time (2008 = 0); (e) EQI for z-score method with uncertainty intervals; (e) EQI for re-scaling method with uncertainty intervals; (f) Z-score EQI change over time (2008 = 0) (Legend: — = 95 <sup>th</sup> percentile; — = 5 <sup>th</sup> percentile) . . . . .	123

5.6	Time series plots (2008 – 2012) for DZ S01006286: (a) EQI for ranking method with uncertainty intervals; (b) Ranked EQI change over time (2008 = 0); (c) EQI for re-scaling method with uncertainty intervals; (d) Re-scaled EQI change over time (2008 = 0); (e) EQI for z-score method with uncertainty intervals; (f) Z-score EQI change over time (2008 = 0) (Legend:  = 95 <sup>th</sup> percentile;  = 5 <sup>th</sup> percentile) . . . . .	124
5.7	Plots showing levels of agreement for EQI from year-to-year (2008 – 2012) . . . . .	126
6.1	Comparing the EQI and the Scottish Index of Multiple Deprivation (Scottish Government, 2012) – contains Ordnance Survey data © Crown copyright and database right (2014) . . . . .	137

# Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Marian Scott for her guidance over the course of this project. I am extremely grateful to the Scottish Environment Protection Agency for funding this research and to Mark Hallard and colleagues at the Environmental and Spatial Informatics Unit for providing direction.

I am grateful to Dr. Duncan Lee, Guowen Huang and Francesca Pannullo for providing some of the data and software used in this project as well as Fiona Fordyce and colleagues at the British Geological Survey for allowing the use of their data.

I would also like to acknowledge Dr. Claire Miller, Dr. Ruth O'Donnell, Dr. Alastair Rushworth, Lauren Sim, Kelly Gallacher and Craig Wilkie for providing advice and help when it was required.

## **Declaration of Originality**

This thesis consists of original work by the author except where clearly stated and referenced to a previously-published work listed in the bibliography.



# Chapter 1

## Background

### 1.1 Introduction

An environmental indicator expresses information about the state of an ecosystem, whether urban or rural (Jørgensen et al., 2005), and is a numerical summary of environmental data. Indicators play an important role in environmental and ecological risk assessment (Suter, 2001) and can also be used as a measure of distance from an environmental goal or target set by decision-makers (Manoliadis, 2002).

The use of indicators in environmental monitoring has been prevalent over the past few decades (Bruno and Cocchi, 2002), with composite indices (sometimes also referred to as composite indicators), comprising of numerous environmental measures, providing a multi-dimensional insight into environmental quality within a region of interest. Nardo et al. (2005) describe a composite index as being the result of a “mathematical combination of individual indicators that represent different dimensions of a concept”. In an environmental context, a composite index is essentially a synthesis of several environmental indicators, leading to a wider and more general picture of the state of an ecosystem.

Composite indices are useful as they are able to communicate the results of statistical analyses in a simple and understandable way to a lay audience. They are able to

summarise complex, multi-dimensional issues, provide information to policymakers and can assess progress, or lack of progress, in tackling an environmental concern over time (Nardo et al., 2005).

However, if poorly constructed or misinterpreted, composite indices may mislead the intended audience or invite overly simplistic policy conclusions (Nardo et al., 2005). Therefore, a transparent and reasoned method of construction is required. Bruno and Cocchi (2002) (derived from Ott (1978)) discuss the possibility of ambiguity (when the global index signals a poor situation when its components suggest otherwise) and eclipsicity (where the global index signals a good situation whilst an assessment the individual indicators provides evidence to the contrary). Therefore, a clear framework for constructing a composite index is essential to avoid the accidental obscuring of the true meaning of the data.

No single method for the construction of composite environmental indices exists (Bruno and Cocchi, 2002) and, hence, numerous efforts have been undertaken in the last few decades to create indices that represent air, water, land and other environmental quality issues, as well as allowing for spatial and temporal comparisons (Hsu et al., 2013).

### **1.1.1 Examples of Past and Current Environmental Indices**

One of the pioneering environmental indices was the Pollution Standards Index that was developed by the United States Environmental Protection Agency (Ott and Hunt, 1976). Ott and Hunt (1976) specified the components of the index, stating that it “includes up to five pollutants . . . and it reports on the maximum pollutant concentration.”

This and subsequent indices have sought to combine environmental monitoring information in order to inform and influence policymakers and the general public (Hsu et al., 2013). The ability to communicate key statistical and scientific findings to

these groups in an accessible and meaningful manner is critically important (Hsu et al., 2013).

An example of a more modern composite index with an environmental component is the Scottish Index of Multiple Deprivation or SIMD (Scottish Government, 2012), which is reported at the resolution of a data zone. Data zones are small-scale geographies used to report many types of official statistics in Scotland. Scotland is divided into 6,505 data zones, each with a population of approximately 800 residents. The Scottish Government (2012) states that data zones are typically small, in a geographical sense, for urban areas and may consist of only a few streets whereas for sparsely-populated rural areas they may cover many square miles. Data zones will be used as the geographical unit to express the results of the composite indices developed in this thesis.

The SIMD allocates a deprivation score to each of Scotland's 6,505 data zones. The overall composite index consists of seven separate domains (or sub-indices) that are widely suspected to contribute to deprivation, namely, employment, income, crime levels, housing, health, education and access to key public services (Scottish Government, 2012). The scores for each of the seven domains are calculated from a number of individual indicators. The domains and the respective indicators that comprise the domains are weighted in order to create the final index.

Eurostat, the Directorate-General of the European Commission responsible for statistics and promoting the harmonisation of methods in producing official statistics across European Union member-states, specified a framework for indicators in the field of environmental protection, in conjunction with the European Environment Agency. As shown in Table 1.1, Eurostat (2014) has defined five types of environmental indicators, covering initial causes of environmental degradation, the subsequent effects of such degradation and eventual responses from policymakers to reverse such effects.

Table 1.1: Indicator framework for environmental protection (Eurostat, 2014)

Indicator	Description	Examples
Driving force indicator	Describes social, demographic or economic developments, the corresponding changes in lifestyles and the overall levels of consumption and production patterns	Population dynamics; GDP
Pressure indicator	Describes developments in the release of substances, physical and biological agents and the uses of land and resources	$CO_2$ emissions; use of natural resources; land use
State indicator	Provides a description of the quantity and quality of biological, chemical and physical variables in certain areas	Global mean temperature; species diversity; atmospheric $CO_2$ concentrations
Impact indicator	Describes the relevance of changes in the state of the environment and implications for ecosystems, human wellbeing and the economy	Percentage of population exposed to noise above thresholds or drinking water below quality standards
Response indicator	Refers to responses by society and policymakers to attempt to prevent, compensate or adapt to changes in the state of the environment	Environmental expenditure; recycling rates

## 1.2 Aspects of Environmental Quality

### 1.2.1 Air

The composition of ambient air in urban environments has been a topic of considerable concern for scientists, policymakers and much of the general public for many decades. The presence of harmful pollutants such as nitrogen dioxide ( $NO_2$ ), carbon monoxide ( $CO$ ), ozone ( $O_3$ ) and particulate matter (usually measured as  $PM_{10}$  or  $PM_{2.5}$ ) in such air is not only detrimental to the environment but also to human health (Guerreiro et al., 2014). Currently, “particulate matter and ozone are Europe’s most problematic pollutants in terms of harm to health” (European

Environment Agency, 2013) and are mostly found in cities, where much of the human population lives. Chiras (2006) states that “natural events such as volcanic eruptions, dust storms and forest fires produce huge quantities of air pollution” on a daily basis, but that “anthropogenic pollutants generally create the most significant long-term threat to the biosphere” (Chiras, 2006). The European Environment Agency (2007) attributes energy consumption, industrial activities, transport demand and agriculture to be the specific anthropogenic forces most directly linked with air quality concerns.

Air pollution is widely regarded as an exacerbating factor in pulmonary health conditions such as asthma and cardiorespiratory disorders with some more serious pollutants such as toxic aromatic hydrocarbons also acting as carcinogens (Hardy, 2003). Hardy (2003) also claims that “every year approximately seven hundred thousand deaths worldwide result from air pollution and the World Health Organisation has ranked air pollution as one of the top ten causes of disability.”

Due to the fact that the majority of anthropogenic air pollution is emitted from highly-populated urban environments, this also results in increased numbers of aforementioned illnesses being contracted due to such vast concentrations of people in these high-risk areas. This could be seen as being something of a ‘vicious cycle’.

Vallero (2014) states that natural “ecosystems can be harmed by numerous human activities that introduce stressors to the environment.” Similarly to human health risks, these stressors may be chemical, physical or biological (Vallero, 2014).

There is a wealth of legislation and government directives regarding the enforcement of measures to reduce air pollution, much of which is decided at a European level (Air Quality in Scotland, 2014). For instance, the 2008 Ambient Air Quality Directive set legally-binding limits for ground-level concentrations of air pollutants (European Environment Agency, 2013), which are only allowed to be exceeded a small number

of times per year.

The first piece of air quality-related legislation enacted in recent U.K. history was the Air Quality Standards Regulations 1989, which implemented EU directives regarding limit values and guide values for sulphur dioxide, suspended particles, lead and nitrogen dioxide (Scottish Environment Protection Agency, 2014). The most important Acts of Parliament in the field of air quality were the Clean Air Act 1993 and the Environment Act 1995, the latter of which established the Environment Agency (in England and Wales) and the Scottish Environment Protection Agency (in Scotland), with both organisations tasked with monitoring environmental issues in their respective territories.

The National Air Quality Strategy (NAQS), published by the then-Department of the Environment (now the Department for Environment, Food and Rural Affairs or DEFRA), established a framework of standards and objectives for the air pollutants of most concern, namely sulphur dioxide ( $SO_2$ ), particulate matter ( $PM_{10}$  and  $PM_{2.5}$ ), nitrogen dioxide ( $NO_2$ ), carbon monoxide ( $CO$ ), lead ( $Pb$ ), benzene ( $C_6H_6$ ) and tropospheric ozone ( $O_3$ ) (Scottish Environment Protection Agency, 2014). The NAQS was updated to its current version in 2007.

The Scottish Parliament has since passed additional legislation to improve air quality standards in Scotland, including the implementation of further EU directives. These directives have set further limits on key ambient air pollutants such as sulphur dioxide, nitrogen dioxide,  $PM_{10}$  and lead. Furthermore, the Air Quality Standards (Scotland) Regulations 2007 introduced detailed requirements for monitoring  $PM_{2.5}$  and stipulated the minimum geographical coverage for the monitoring and dissemination of information to the general public (Scottish Environment Protection Agency, 2014).

### 1.2.2 Water

Nesaratnam (2014) states that “to some extent, a river is a self-renewing resource”, meaning that when pollutant levels are low and intermittent, a river is able to discharge any harmful pollutants (Nesaratnam, 2014). This process is often aided by organic bacteria present in the water. However, some pollutants, either in low or high concentrations, can kill off bacteria and prevent the river from “self-purifying” (Nesaratnam, 2014). For instance, arsenic contamination of water, which has concerned public health scientists in many parts of the world (Ahuja, 2013), is a particularly serious problem.

In many industrial areas, these contamination issues are of particular concern due to the harmful potential of industrial effluents (Nesaratnam, 2014). In urban areas, surface water run-off is also an acknowledged threat to water quality, as these river contaminants often contain chemicals such as lead, from motor vehicles.

As with legislation and regulations relating to air quality, the European Union is the principal driving force behind the creation and enforcement of water quality standards. In 2003, the Water Framework Directive or WFD (European Union, 2000) imposed wide-ranging responsibilities on European Union member-states to monitor water quality and related issues in their respective nations, often requiring domestic legislative changes (Benedini and Tsakiris, 2013).

In Scotland, the WFD led to the enactment of the Water Environment and Water Services (Scotland) Act 2003, which gave Scottish Ministers powers to introduce regulatory controls over wetlands, rivers, lochs, estuaries (termed transitional waters), coastal waters and groundwater (Scottish Environment Protection Agency, 2014). Subsequent EU and Scottish legislation has been pursued in this field.

Benedini and Tsakiris (2013) outline how the aims of the WFD are achieved:

- Preventing further deterioration of water quality
- Promoting sustainable water use
- Conserving aquatic environments through reduction of discharges, emissions and losses of priority substances
- Reducing the contamination of groundwater by pollutants
- Mitigating the effects of floods and droughts

### **1.2.3 Land**

Garrigues et al. (2012) state that “soils are an essential resource in both managed and natural systems and maintaining soil quality is critical to sustainable development.” Soil quality is the ability of a particular soil to engage in natural environmental processes such as facilitating animal and plant life and positively contributing to human health and air and water quality (Karlen et al., 1997; Garrigues et al., 2012).

The Scottish Soil Framework (Scottish Government, 2009) outlines the following seven key roles that soil plays in the environment:

- Provides the basis for food and biomass production
- Controls and regulates environmental interactions such as water flow and quality
- Stores carbon and maintains the gaseous balance of ambient air
- Provides animal habitats and sustains biodiversity
- Provides platforms for buildings, roads and other structures
- Provides raw materials
- Preserves cultural and archaeological heritage

Soil quality is at risk from a number of natural processes and anthropogenic threats such as climate change, land use change and land management practices (Scottish Environment Protection Agency, 2014). It is incredibly difficult, if not impossible, to



restore soil quality and, in some cases, the soils themselves may become pollutants, infecting local watercourses for instance (Scottish Environment Protection Agency, 2014). In cities such as Glasgow, past wide-ranging industrial practices were the principal driver of land and soil contamination (Scottish Environment Protection Agency, 2014).

Unlike air and water quality, relatively few European Union member-states have specific legislation relating to soil quality and protection. In 2006, the European Commission adopted a ‘Thematic Strategy for Soil Protection’ that identified threats to soil quality and outlined proposals for a Soil Framework Directive similar to the Water Framework Directive (Dobbie et al., 2011). However, member-states did not reach agreement on the legislation and it was withdrawn in May 2014 (European Commission, 2014). This means that soil quality still lacks the same strict legal protection as air and water quality, at least at European level.

However, in Scotland the Scottish Government established the Scottish Soil Framework, recognising that soil was a vital part of the economy, environment and heritage of the nation (Scottish Government, 2009).

### **1.3 Construction of Composite Indices**

Despite the absence of a common approach to developing composite indices, as mentioned in Section 1.1, a number of methods have been developed in both environmental and non-environmental contexts to create such indices.

As also discussed in Section 1.1, composite indices can be somewhat misleading if poorly constructed and, hence, a degree of scepticism exists amongst some within the statistical community around the issue of composite indices as a whole (Saisana et al., 2005). Saisana et al. (2005) attributes this to a feeling that large amounts of work in data collection and analysis are obscured by its simplification into a single

number or index.

However, Saisana et al. (2005) acknowledges the view that the simplicity of composite indices is much more helpful to non-statisticians and the general public, and this alone justifies their development, provided that a clear, transparent and statistically-sound methodology is specified and followed. To assist in this process, the Organisation for Economic Co-operation and Development (OECD), in collaboration with the Joint Research Centre of the European Commission, has published a list of ten key steps that should be considered when creating a composite index, shown in Table 1.2.

Table 1.2: Steps for constructing a composite index (OECD, 2008)

Step	Notes
(1) Theoretical Framework	Provides basis for the selection and combination of variables under a fitness-for-purpose principle
(2) Data Selection	Data should be selected on analytical soundness, measurability and relevance of the indicators
(3) Imputation of Missing Data	Often needed in order to provide a complete data set
(4) Multivariate Analysis	Should be used to study overall structure of the data set and guide methodological choices (e.g. weighting, aggregation)
(5) Normalisation	Should be carried out to render the variables comparable
(6) Weighting and Aggregation	Should be done along the lines of the underlying theoretical framework
(7) Uncertainty and Sensitivity Analysis	Should be undertaken to assess the robustness of the composite index
(8) Return to Data	Reveals the main drivers for an overall good or bad index performance
(9) Links to Other Indices	Should be made to correlate the composite index with existing comparable indices
(10) Visualisation of the Results	Should be undertaken to enhance interpretability

This section will discuss literature that contain examples of the development of both indicators and composite indices.

A typical example of using statistical methods to create an index for air quality is that of Bruno and Cocchi (2002). The method involves using elementary data  $X_{ijh}$ , where  $i = 1, \dots, I$  indexes the air quality monitoring sites,  $j = 1, \dots, J$  indexes the pollutants being measured at each site and  $h = 1, \dots, H$  indexes time occurrences, with equal spacing between all time points. These data are organised into a three-dimensional matrix. The units of measurement of the pollutant values are not necessarily the same as different pollutants may be measured and reported on different scales. This is accounted for as part of this procedure through normalisation of variables.

Nardo et al. (2005) list the most common methods of normalisation, such as ranking, standardisation (including z-scores), re-scaling to the  $[0, 1]$  range, distance to a reference point, categorical scales and cyclical indicators, amongst others. They also note that in analyses involving extreme values, normalisation methods that are based on the standard deviation are preferable.

Returning to the index construction method employed by Bruno and Cocchi (2002), what follows is a process of aggregation where a function  $q$  is applied to the time component,  $h$ , of the elementary data to create a time synthesis i.e.

$$X_{q\{ij\}} = q(X_{ijh}) \tag{1.1}$$

The application of the function  $q$  to the elementary data generates an  $I \times J$  matrix, which contains, for the  $i^{th}$  monitoring site, the time synthesis for each pollutant and each column contains, for the  $j^{th}$  pollutant, the time synthesis of each site.

For subsequent aggregation procedures, it is necessary to use a form of standard-

isation due to the potentially differing units of measurement used to report the pollutant concentrations, as aforementioned. Bruno and Cocchi (2002) suggest normalising the various pollutants by utilising a ratio of each pollutant concentration by a standard concentration value. However, they also observe that when a continuous scale is used, the segmented line function proposed by Ott and Hunt (1976) for the Pollution Standards Index (discussed in Section 1.1.1) is preferable:

$$f_R(Y) = \frac{b_{c+1} - b_c}{a_{(c+1)j} - a_{cj}}(Y - a_{cj}) + b_c \quad (1.2)$$

where  $a_{cj} < Y < a_{(c+1)j}$ ,  $c = 1, \dots, C \ \forall \ j = 1, \dots, J$ . Bruno and Cocchi (2002 – derived from Ott and Hunt (1976)) define the above parameters:  $Y$  is the pollutant concentration value obtained by any previous aggregation steps, the  $a_{cj}$ 's represent the thresholds for different air quality classes in terms of health risk and the  $b_c$ 's are the standardised thresholds for each pollutant.

The choice of whether to aggregate by site then pollutant or pollutant then site can result in different values for the final index. Here, the aggregation was performed by site and then by pollutant so that the standardising of the pollutants could be delayed.

To eliminate the spatial dimension, under the assumption that the monitoring sites are homogeneous, a second function  $g$  is applied to the index from Equation 1.1 (Bruno and Cocchi, 2002) i.e.

$$X_{g\{j\}} = g(X_{q\{ij\}}) \ \forall \ j = 1, \dots, J. \quad (1.3)$$

This creates a  $J$ -dimensional vector, the components of which require standardisation as aforementioned in order to eliminate the units of measure. A  $g$ -type function,  $g^*$ , is applied to the standardised pollutant values to obtain a final single value (Bruno and Cocchi, 2002):

$$I_{g^*(g)} = g^* f(X_{\{j\}}) \quad (1.4)$$

Bruno and Cocchi (2002) proposed using either the median value or maximum value for each pollutant for the aggregation function  $g$ , as both produce good results even in the presence of missing data and the median is also a robust statistic (i.e. not unduly influenced by extreme values).

Bruno and Cocchi (2002) observe that “a common critic to the proposal of synthetic values is in fact that they do not account for any form of variability”. They attempt to address this by including measures of dispersion, which are obtained by calculating ratios between the final index where  $g^*$  is the median and the final index where  $g^*$  is the maximum. This can be achieved regardless of whether space or pollutants are aggregated into the index first.

The method employed by Bruno and Cocchi (2002) was developed by Lee et al. (2011) who incorporated a pre-aggregation stage involving the geostatistical modelling of available data in order to predict air quality concentrations at unobserved locations. This was achieved by the use of a Bayesian geostatistical model and then implementing the spatial and pollutant aggregation methods developed by Bruno and Cocchi (2002). This procedure allowed for the construction of credible intervals for the spatial aggregation stage (Lee et al., 2011), meaning that a range of plausible values for each indicator was available as opposed to one single value.

Lee et al. (2011) also considered the possibility of ‘preferential sampling’, where the choice of sampling locations is not stochastically-independent (Diggle et al., 2010) and does not allow for spatial representativeness (Lee et al., 2011). Diggle et al. (2010) states that preferential sampling can result in bias with regards to parameter estimation and spatial prediction. Lee et al. (2011) accounted for these possibilities by first fitting a geostatistical model assuming independence and another allowing

for preferential sampling, which could also be considered as a form of sensitivity analysis.

An example of a composite environmental quality index which combines an air quality indicator with that of another environmental quality determinand was developed by Silva and Mendes (2012), who sought to quantify the extent to which air and noise pollution objectives were met in Viana do Castelo, Portugal. The index consisted of a weighted linear combination of two individual, normalised sub-indices or domains (for air pollution and noise pollution), both of which were allocated equal weights. The final index was a numerical value ranging from 0 to 1.

The air pollution sub-index was specified as follows (Silva and Mendes, 2012):

$$\text{Air} = \sum_i w_i c_i \times \prod_i v_i \quad (1.5)$$

where  $w_i$  is the relative weight of pollutant  $i$ ,  $c_i$  is the normalised concentration of pollutant  $i$  and  $v_i$  is an indicator variable of the legal limit violation  $L_i$  of pollutant  $i$ , i.e.

$$\begin{cases} v_i = 1 & \text{when } c_i \leq L_i \\ v_i = 0 & \text{when } c_i > L_i \end{cases} \quad (1.6)$$

The air index consisted of five pollutants, carbon monoxide ( $CO$ ), nitrogen dioxide ( $NO_2$ ), particulate matter (measured as  $PM_{10}$ ), benzene ( $C_6H_6$ ) and ozone ( $O_3$ ), all of which received an equal weighting of 0.2.

However, because the pollutants were measured on different scales, they were standardised in order to transform the scale to a normalised range (i.e. between 0 and 1). Silva and Mendes (2012) used a sigmoidal function for this step i.e.

$$\text{Score} = \frac{1}{\sin^2 \alpha} = \cos^2 \alpha \quad (1.7)$$

where

$$\alpha = \frac{(x - x_a)}{(x_b - x_a)} \times \frac{\pi}{2} \quad (1.8)$$

where  $x$  is the concentration value being normalised and  $x_a$  and  $x_b$  are control points in the function, which are specified further in Silva and Mendes (2012). The formulation of the noise index is also not detailed here as it is not relevant to this thesis but is defined in Silva and Mendes (2012).

Some composite indices consist of dozens, if not hundreds, of indicators, meaning that a clearly-specified construction methodology (following, for instance, a framework like that in Table 1.2) is even more critical so that the index is transparent and open to scrutiny. Developers of composite indices must consider such details as indicator and domain weightings, which can be complex to define and can have an impact on the final index. However, sensitivity and uncertainty analyses should quantify such impacts, if any.

An important composite index in Scotland is the aforementioned Scottish Index of Multiple Deprivation (SIMD) which, as described in Section 1.1.1, is developed at a small-scale geographical level (a data zone) and seeks to identify those places in Scotland suffering from deprivation (Scottish Government, 2012).

The SIMD is comprised of thirty-eight indicators across seven domains or sub-indices (employment, income, health, education, geographic access to key public services, crime and housing) and is reported on a ranked scale from 1 (the most deprived data zone) to 6,505 (the least deprived data zone). The domains (and in some cases, the

individual indicators) require weighting as some domains may be more important than others in measuring deprivation.

The criteria for selecting an indicator were as follows (Scottish Government, 2012):

- The indicator should be domain-specific and appropriate for the purpose of measuring deprivation
- The data should be as recent as possible
- The data set in question should be capable of being updated on a regular basis
- The data should be statistically-robust
- The indicator should measure major features of a given type of deprivation (not conditions experienced by relatively few individuals or areas)

However, not all data zone ranks are calculated in the same way, with some being straightforward counts (income, employment, housing and crime) and the remaining domains (health, education and access to services) being constructed from weighted indicator scores (Scottish Government, 2012).

Figure 1.1 details the weighting and aggregation methods used to create the final deprivation index.

However, as is common in the construction of many composite indices, prior to the weighting and aggregation of individual indicators or domains, they must be normalised. The domains within the SIMD are normalised by ranking the scores (Scottish Government, 2012). The ranks are then exponentially-transformed to avoid high ranks in one domain effectively ‘cancelling out’ low ranks in another.

As aforementioned, the final SIMD results are then ranked from 1 (the most deprived data zone in Scotland) to 6,505 (the least deprived) in order to create the final index. It should be noted that the ranks of the data zones can only be compared relative



## SIMD 2012 Methodology

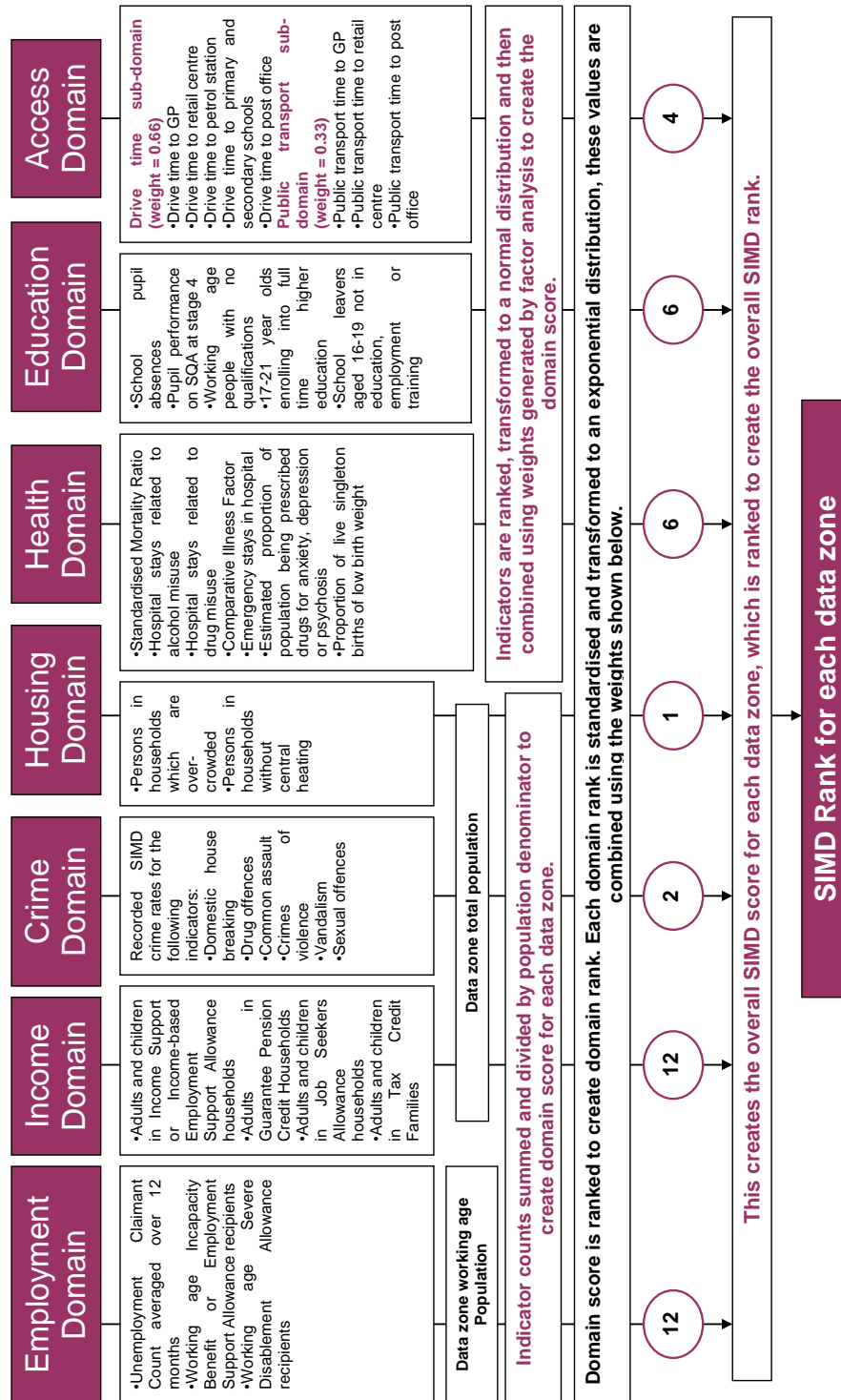


Figure 1.1: Indicators, aggregation methods and weightings used by the Scottish Index of Multiple Deprivation (Scottish Government, 2012)

to each other. For instance, a data zone which is ranked 50<sup>th</sup> is not necessarily twice as deprived as the data zone which is ranked 100<sup>th</sup>. All that can be inferred from the overall index is that the data zone which is ranked 50<sup>th</sup> is more deprived than the data zone which is ranked 100<sup>th</sup>.

A similar approach was taken by Richardson et al. (2013), who created a composite index for multiple deprivation for each data zone in the Scottish local authority of South Lanarkshire, although with a more health-based focus. The index consisted of three domains or sub-indices (hazardous environmental indicators, undesirable environmental indicators and salutogenic environmental indicators).

Richardson et al. (2013) followed a clearly-defined strategy for the creation of the composite index, consisting of the following steps:

- Selection of indicators
- Selection of geographical unit
- Constructing and testing the domains
- Constructing the index

### **Selection of indicators**

The indicators which comprised each domain were air pollution, noise pollution and traffic environment for the hazardous environments domain, undesirable land uses and crime rates in the undesirable environments domain and availability of urban green space, public transport, paths and access to key services for the salutogenic environments domain.

### **Selection of geographical unit**

As aforementioned, the geographical unit of analysis was the data zone, like that used by the SIMD (Scottish Government, 2012).

### **Constructing and testing the domains**

Richardson et al. (2013) state that the “construction of the domains was informed by the methodology used in the development of indexes of multiple deprivation in the United Kingdom”, such as SIMD (Scottish Government, 2012). Each data zone in South Lanarkshire was allocated a proportional rank for each of the nine indicators, ranging from 0 for the data zone with the best indicator value to 1 for the data zone with the worst indicator value.

Like SIMD (Scottish Government, 2012), the indicators were then exponentially-transformed in order to emphasise the highest values and avoid high values of one indicator cancelling out the low values of another (Richardson et al., 2013). Afterwards, the transformed indicator scores for each data zone were summed to produce scores for each of the three domains (Richardson et al., 2013).

However, after testing for valences, the positive or negative direction of the domains (Messer et al., 2014), it was discovered that the direction of the salutogenic environment domain was the opposite of what was expected or rational (i.e. better salutogenic conditions were leading to more environmental deprivation) and, hence, it was concluded that the domain was not fit for purpose and it and its component indicators were excluded from the final index (Richardson et al., 2013).

### **Constructing the index**

The five remaining indicators (three from the hazardous environments domain and two from the undesirable environments domain, respectively) were then combined to form a single index using factor analysis. Richardson et al. (2013) state that “the technique is based on the premise that the indicators that are most highly correlated with the domain will also be highly correlated with each other”.

The indicators were first ranked and transformed to a Gaussian distribution with the subsequent analysis identifying a single factor that explained the variability better

than the original indicators themselves (i.e. an eigenvalue greater than 1). The condition index of the overall factor analysis was 2.2, which led to the conclusion that there was little evidence of multicollinearity between the indicators present (Richardson et al., 2013). Finally, the indicators were multiplied by the loadings (or indicator weights in this context) for the first factor and summed to produce an index score for each data zone in South Lanarkshire.

Richardson et al. (2013) observe that the final index, termed the South Lanarkshire Index of Multiple Environmental Deprivation, was normally-distributed, ranged from -4.0 (least environment deprivation) to +4.2 (greatest environmental deprivation) and the mean value of the index was zero. As a last act, they also partitioned the data zones into quintiles to represent increasing environmental deprivation.

In an international context, the Environmental Performance Index (Hsu et al., 2014), created jointly by the Centre for Environmental Law and Policy at Yale University and the Centre for International Earth Science Information Network at Columbia University, is a composite index combining indices relating to nine ‘issues’ (comparable to domains or sub-indices) and nineteen ‘indicators’, shown in Table 1.3.

The Environmental Performance Index (Hsu et al., 2014) states that “each indicator is weighted within each issue category to create a single issue category score.” Each indicator is usually considered equal to other indicators within the same issue category and is weighted appropriately but this depends on the quality of the underlying data set and on the relevance or fit of the indicator (Hsu et al., 2014). Hence, in some cases an indicator will be weighted less heavily. The Environmental Performance Index uses the criteria shown in Table 1.4 to decide whether an indicator should qualify for the maximum possible weighting (Hsu et al., 2014).

Through the nine ‘issues’ and the nineteen ‘indicators’ listed in Table 1.3, the Environmental Performance Index allocates each participating country a score based

Table 1.3: Structure of the Environmental Performance Index (Hsu et al., 2014)

Issue Categories	Indicators
Health Impacts	Child Mortality
Air Quality	Household Air Quality; Average Exposure to $PM_{2.5}$ ; $PM_{2.5}$ Exceedance Level
Water and Sanitation	Access to Drinking Water; Access to Sanitation
Water Resources	Wastewater Treatment
Agriculture	Agricultural Subsidies; Pesticide Regulation
Fisheries	Coastal Shelf Fishing Pressure; Fish Stock
Forests	Change in Forest Cover
Biodiversity and Habitat	Natural Biome Protection; Global Biome Protection; Marine Protected Areas; Critical Habitat Protection
Climate and Energy	Trend in $CO_2$ Emissions per kWh; Change of Trend in Carbon Intensity; Trend in Carbon Intensity

Table 1.4: Selection criteria for data in the Environmental Performance Index (Hsu et al., 2014)

Criterion	Notes
Relevance	Indicator tracks the environmental issue in a manner that is applicable to countries under a wide range of circumstances
Performance orientation	Indicators provide empirical data on ambient conditions or on-the-ground results for the issue of concern, or is it a ‘best available data’ proxy for such outcome measures
Established scientific methodology	Indicator is based on peer-reviewed scientific data or data from the United Nations or other institutions charged with data collection
Data quality	Data represent the best measure available; all potential data sets are reviewed for quality and verifiability; those that do not meet baseline quality standards are discarded
Time series availability	Data have been consistently measured across time and there are ongoing efforts to continue consistent measurement in the future
Completeness	Data set needs to have adequate global and temporal coverage to be considered

on relevant environmental concerns. The presentation of one final number from an index should give each country a clear indication as to the state of its environment. Repeated analyses over time should also allow comparisons with previous performance and show whether a country's environmental condition has improved or deteriorated, in a relative sense.

A composite index that is developed according to a completely different approach are the indices developed by the Canadian Council of Ministers of the Environment (CCME), an intergovernmental body that consists of environment ministers from the federal, provincial and territorial governments in Canada. There are composite indices measuring both water and soil quality.

The CCME Water Quality Index (CCME, 2001) consists of three factors: scope, frequency and amplitude. It also requires thresholds or limits for each determinand to be known, as this composite index is used to quantify if, and by how much, concentrations of determinands have breached their specified limits.

The scope factor, known as  $F_1$ , represents water quality guideline non-compliance over the time series i.e.

$$F_1 = \left( \frac{\text{Number of failed variables}}{\text{Total number of variables}} \right) \times 100 \quad (1.9)$$

The frequency factor, known as  $F_2$ , represents the percentage of individual tests that do not meet the objectives i.e.

$$F_2 = \left( \frac{\text{Number of failed tests}}{\text{Total number of tests}} \right) \times 100 \quad (1.10)$$

The amplitude factor,  $F_3$  (defined shortly), represents the amount by which failed test values do not meet their objectives. This is a multi-stage calculation.

### Stage 1

It is necessary to calculate the number of times where individual concentrations are greater than or less than (depending upon the direction of the limit) the specified threshold, termed an ‘excursion’. If the test value must not exceed the objective:

$$\text{Excursion}_i = \left( \frac{\text{Failed Test Value}_i}{\text{Objective}_j} \right) - 1 \quad (1.11)$$

Conversely, if the test value must not fall below the threshold:

$$\text{Excursion}_i = \left( \frac{\text{Objective}_j}{\text{Failed Test Value}_i} \right) - 1 \quad (1.12)$$

### Stage 2

The overall amount by which individual tests are non-compliant is calculated by summing the excursions of individual tests from their specified thresholds and dividing by the total number of tests. This is known as the normalised sum of excursions:

$$\text{nse} = \frac{\sum_{i=1}^n \text{Excursions}_i}{\text{Number of tests}} \quad (1.13)$$

### Stage 3

Finally, the amplitude factor,  $F_3$ , is calculated by scaling the normalised sum of the excursions from their objectives (nse) using an asymptotic function:

$$F_3 = \left( \frac{\text{nse}}{0.01 \times \text{nse} + 0.01} \right) \quad (1.14)$$

The final CCME WQI is then calculated:

$$\text{CCME WQI} = 100 - \left( \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \right) \quad (1.15)$$

The scale factor of 1.732 is derived from the fact that the individual index factors have a maximum of 100 i.e. the vector length can reach a maximum of:

$$\sqrt{100^2 + 100^2 + 100^2} = \sqrt{30000} = 173.2 \quad (1.16)$$

Dividing by 1.732 constrains the vector to a maximum of 100. This results in the following ordinal scale:

- Excellent ( $95 \geq \text{CCME WQI} \leq 100$ ) – all measurements are within objectives virtually all of the time
- Good ( $80 \geq \text{CCME WQI} \leq 94$ ) – most measurements within objectives
- Fair ( $65 \geq \text{CCME WQI} \leq 79$ ) – some departures from objectives
- Marginal ( $45 \geq \text{CCME WQI} \leq 64$ ) – frequent departures from objectives
- Poor ( $0 \geq \text{CCME WQI} \leq 44$ ) – very frequent departures from objectives or virtually no meetings of objectives at all

The CCME Soil Quality Index (CCME, 2007) is calculated in a very similar manner to the WQI.

One criticism of the CCME Water Quality Index (CCME, 2001) is that its indicators are all equally-weighted, which may not be desirable or sensible. For instance, chemicals which are poisonous in high quantities (such as arsenic) should be more heavily-weighted in the final index construction than a less important indicator. Mohebbi et al. (2013) accounted for this by allowing for unequal indicator weighting, although the index construction methodology otherwise remained similar.

Finally, a composite environmental index with great relevance to the methodology used in this thesis, albeit with a public health aim for the United States, is that created by Messer et al. (2014), which involves the weighted aggregation of several domains quantifying different aspects of the overall environment. The results were also presented and visualised for distinct, non-overlapping areas, which is also an



aim of this thesis. The five domains were air, water, land, socio-demographic issues and the built environment (the first three of which are used in this thesis; the latter two are not covered here). The geographic level of analysis was the U.S. county, which is the third-level administrative tier below federal and state authorities.

Important pre-analysis data issues included variable multicollinearity and variable missingness. For the former, correlation coefficients were calculated within each of the five domains and where correlation that exceeded 0.7 was present between two variables, only one was selected for inclusion in the final index (Messer et al., 2014). The choice of which variable to exclude was usually based on the proportion of missing values, where the variable with the lowest proportion was the one that was selected (Messer et al., 2014).

To tackle variable missingness, a decision was made as to whether missing data were actually missing or instead represented true zero values. For instance, when crime data were missing it was considered to be truly missing as crime would naturally occur in every county, but for beach closure data, true zeros were considered not to be missing as not all counties are in coastal areas. If data for a variable were missing for more than fifty percent of counties, then the variable was excluded from the index (Messer et al., 2014).

Also, domain variables were log-transformed where appropriate to better satisfy normality assumptions. For variables which contained true zeros, the log-transformation was achieved by first adding a constant (half of the non-zero minimum value) to all observations.

For the air domain, daily concentrations for six indicator pollutants were temporally averaged to ascertain annual mean concentrations for each monitoring station from 2000 to 2005. These annual means were then subject to spatiotemporal kriging in order to estimate annual concentrations at each geographical county centroid. An

exponential covariance structure was used for the spatial covariance. The water and soil domains were created using a similar procedure (Messer et al., 2014).

As with Richardson et al. (2013), a final important step in data preparation was determining valences for each indicator and domain, checking whether the effect of each component is positive or negative with regards to environmental quality (Messer et al., 2014). The higher values of domains or indicators which are known to negatively affect environmental quality should represent poorer quality; likewise, higher values of components known to positively contribute to environmental quality should represent better overall environmental quality (Messer et al., 2014).

The variables from each domain were aggregated to create a domain-specific county-level data set. Principal components analysis was used to combine the individual domains using the variable loadings (or domain weights). The respective loadings for each variable were then multiplied by the mean value for each county and the weighted means were summed. Only the first principal component was retained by Messer et al. (2014), as it accounted for the largest proportion of the total variability. This first principal component was also normalised to have a mean of 0 and a standard deviation of 1, achieved by dividing the index by the square of the eigenvalue (derived from Kim and Mueller, 1978).

A further analysis undertaken by Messer et al. (2014) was to stratify the final indices according to the “rural-urban continuum”, meaning that the index would be more useful in their circumstances if it allowed for additional information on whether the county was primarily urban or rural. At the stage of calculating principal components, the analysis was conducted for each of the rural-urban strata.

## 1.4 Overview and Aims of Thesis

The overall aim of this thesis is to investigate and develop methods for constructing composite indices for environmental quality at a small spatial scale, with environmental quality defined as being composed of air, soil and water quality. It is also of interest to quantify the level of uncertainty within such indices.

In this thesis, a composite environmental quality index will be created for the Greater Glasgow region at data zone level, incorporating information about air, water and soil pollution levels. The final index should be comparable with other data zone-level indices used by the Scottish Government and other public bodies in Scotland. The composite index will initially be created for the year 2011 only and then compared with similarly-constructed indices for other years.

Figure 1.2 shows the region of Greater Glasgow for which the composite index will be constructed, comprising of 1,748 data zones across seven local authorities. The study region was limited to a spatial domain in which enough air, water and soil data were available and fit for use. This was most important for the soil data as predicting outwith the vicinity of the data is unreliable. Consequently, the far north-west of West Dunbartonshire was not included as was the eastern half of North Lanarkshire and the rural south of South Lanarkshire. The local authorities of Glasgow City, East Dunbartonshire, Renfrewshire and East Renfrewshire were included in their entirety.

Chapter 2 of this thesis will elaborate on the data used in this project and carry out spatial and temporal exploratory analyses which will inform subsequent processes.

Chapter 3 will concentrate on modelling the data so that pollutant concentration predictions can be made in every data zone across the study region. These predicted data can then be used to construct composite environmental indices for Greater

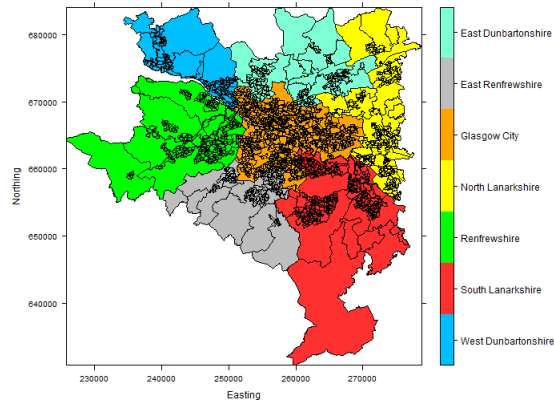


Figure 1.2: Map of study region with local authority and data zone boundaries – contains Ordnance Survey data © Crown copyright and database right (2014)

Glasgow.

Chapter 4 will involve the construction of an environmental quality index (EQI) for Greater Glasgow for the year 2011, drawing on the examples discussed in Section 1.3. It will focus on three different normalisation methods that result in different index scales but ultimately measure the same phenomenon.

Chapter 5 will assess the EQI created in Chapter 4 by undertaking an uncertainty analysis of the results and then using the indices to check how environmental quality varies over a period of five years from 2008 to 2012.

Chapter 6 will discuss the main results shown in this thesis and make final conclusions with regards to the aims of the project.

# Chapter 2

## Data and Exploratory Analysis

This chapter will introduce the data that will be used to construct the composite indices later in this thesis. An exploratory analysis of the data will also be performed in order to identify key spatial and temporal patterns.

### 2.1 Air

The air quality data used in this thesis were collected from three sources. Firstly, data from local monitoring stations across the study region were downloaded from the Air Quality in Scotland database (<http://www.scottishairquality.co.uk>), which is maintained by the Scottish Government. Secondly, data from automatic  $NO_2$  diffusion tubes were provided by local authorities. Lastly, 1 km by 1 km grids consisting of a pollution estimate from an atmospheric dispersion model were downloaded from the website of the Department for Environment, Food and Rural Affairs or DEFRA (<http://www.uk-air.defra.co.uk>).

#### 2.1.1 Air Pollutants

The air quality component of this thesis focussed on four pollutants that are prevalent in ambient urban air and are known to have negative effects on air quality. The unit of measurement for each pollutant was micrograms per cubic metre ( $\mu\text{g}/\text{m}^3$ ).

### **Nitric Oxide ( $NO_X$ ) and Nitrogen Dioxide ( $NO_2$ )**

Nitrogen molecules account for approximately 79% of the Earth's atmosphere (Vallero, 2014).  $NO_2$  is the  $NO_X$  compound that is the greatest concern to human health, as even short-term exposure can lead to respiratory effects.

### **Particulate Matter – $PM_{2.5}$ and $PM_{10}$**

Particulate matter refers to a mixture of both solid and liquid particles suspended in the air with a wide range of sizes and chemical compositions (European Environment Agency, 2013). The numerical subscripts 2.5 and 10 refer to the maximum size of the particles in micrometres. They can be formed from primary sources (e.g. chimneys) or secondary sources through oxidation and transformation of primary gaseous emissions (European Environment Agency, 2013). Studies have attributed the most severe health effects from air pollution to particulate matter, such as in the respiratory, cardiovascular, immune and neural systems (European Environment Agency, 2013).

Data were available for other pollutants such as carbon monoxide ( $CO$ ), ozone ( $O_3$ ) and sulphur dioxide ( $SO_2$ ) but were monitored at very few spatial locations. Also, the aforementioned DEFRA local authority modelled grids were only available for the four selected pollutants. It was decided that only pollutants that had been monitored at no less than fifty spatial locations would be included, given the large geographical extent of the study region.

### **2.1.2 Air Quality Monitoring Station Data**

Figure 2.1(a) displays the air monitoring stations that have been active for at least some time since the year 2000 and before early 2014 and have collected data on at least one of  $NO_2$ ,  $NO_X$ ,  $PM_{10}$  or  $PM_{2.5}$ . The stations cover the local authorities of Glasgow City, East Dunbartonshire, West Dunbartonshire, Renfrewshire, East Renfrewshire, North Lanarkshire and South Lanarkshire. Figure 2.1(b) shows the locations of monitoring stations within the local authority of Glasgow City only.

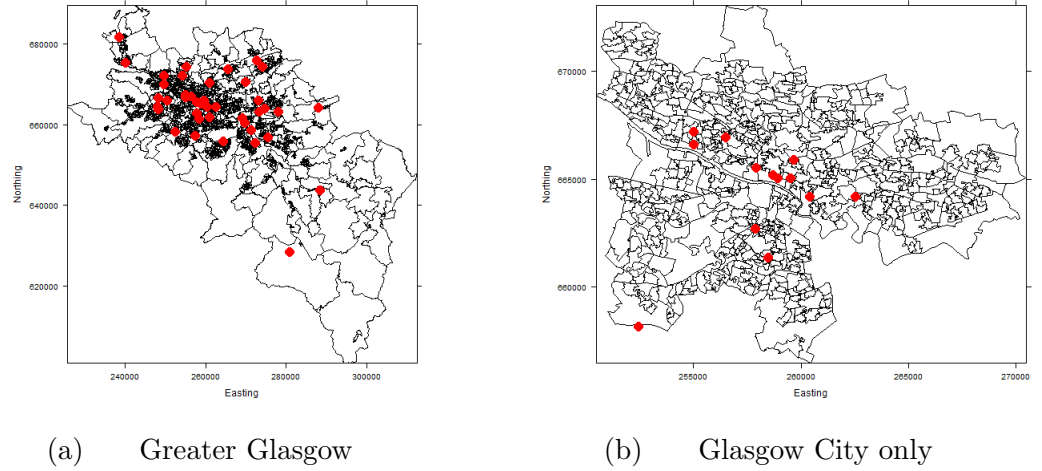


Figure 2.1: Maps of air monitoring stations – contains Ordnance Survey data

© Crown copyright and database right (2014)

Each air monitoring station is classified as ‘roadside’, ‘kerbside’, ‘urban background’ or ‘rural’ depending upon its proximity to a road and, providing it is not a rural station, its location in the street.

Figure 2.1(a) shows that the monitoring stations are largely clustered around the highly-urbanised area in and around the city of Glasgow with very few locations in peripheral areas. In Figure 2.1(b), even the monitoring stations within the city of Glasgow are largely clustered in the most highly-populated areas such as the city centre and the adjoining ‘West End’. The obvious exception is the monitoring station situated in the far south-west of the city.

The exploratory analysis for the air monitoring station data for each of the four pollutants involved largely similar procedures. Only that for nitrogen dioxide ( $NO_2$ ) is detailed here with the corresponding analyses for remaining three pollutants being displayed in Appendix A. Also, the exploratory analysis has been restricted here to six selected monitoring stations, three of which were located in Glasgow City and three of which were located in other local authorities. The exception is  $PM_{2.5}$ , which was only measured at three monitoring stations and its exploratory analysis consists

of results from all three stations.

## Exploratory Analysis

Table 2.1 shows descriptive statistics for six selected monitoring stations for the year 2011.

Table 2.1: Descriptive statistics for selected  $NO_2$  air quality monitoring stations in 2011 (units of measurement =  $\mu\text{g}/\text{m}^3$ )

Monitoring Station	Mean	St. Dev.	Min	Q1	Median	Q3	Max
Glasgow Byres Road	41.54	23.38	2.00	25.00	38.00	55.00	180.00
Glasgow Centre	34.48	20.37	0.00	19.00	31.00	46.00	149.00
Glasgow Waulkmillglen Reservoir	10.52	14.69	0.00	1.90	3.80	11.50	118.40
Bearsden	39.47	28.21	0.00	19.00	34.00	52.00	195.00
Clydebank	20.71	20.44	0.00	6.00	13.00	29.00	136.00
East Kilbride	40.66	24.82	0.00	23.00	36.00	55.00	281.00

Since air pollution concentrations should be non-negative, a log-transformation may be appropriate, as in many environmental settings. The validity of the transformation of  $NO_2$  concentration was investigated using a histogram of the data. Figure 2.2(b) shows that such a log-transformation largely fails to produce a symmetric, bell-shaped distribution as a heavy tail remains. Such a distribution may help in any statistical modelling under a normality assumption. However, the log-transformed data are more bell-shaped than the raw data and, hence, the transformation may still be useful in any modelling process.

The proportion of zero values in the data was small, since air pollution is almost certain to be present in cities such as Glasgow and the surrounding area. Also, zero values may be the result of a limit of detection issue rather than measurements



being genuinely zero. Hence, they were excluded from the data set prior to any log-transformation.

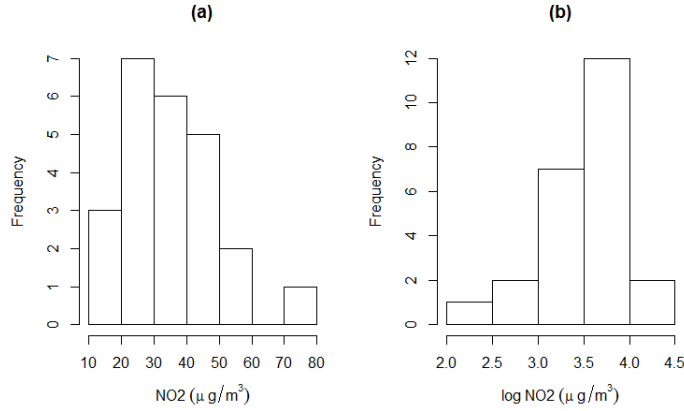


Figure 2.2: (a) histogram of  $NO_2$  concentration ( $\mu g/m^3$ ) values; (b) histogram of log-transformed  $NO_2$  concentration ( $\mu g/m^3$ ) values

Figures 2.3 through 2.6 show how the pollutant concentration varies on annual, monthly, daily (within week) and hourly scales, respectively. Day-within-month plots were also constructed but did not show much variability across one month, with a day-within-week scale being considered more informative (i.e. regarding weekdays versus weekends).

Figure 2.3 displays time series plots for the log-transformed  $NO_2$  concentrations for the six chosen monitoring stations. It is difficult to infer any long-term trends from the plots, suggesting that  $NO_2$  concentrations may have remained relatively stable over the length of the time series. However, seasonal patterns are clearly visible at most locations. At Byres Road, for instance, it appears as if the  $NO_2$  concentration level increases and decreases quite regularly. The occasional periods of missing data should also be noted, shown by gaps in the time series plots.

Figure 2.4 shows that the log-transformed  $NO_2$  concentration level appears to be at its peak in the winter months and is lower in the warmer months. Most of the

boxplots also show a smaller degree of variability in these concentrations in the summer as the size of the boxes is generally smaller. Figure 2.5 shows that there does appear to be variability in  $NO_2$  concentration across one week, with weekdays registering higher levels. Saturdays and Sundays exhibit lower concentrations.

Figure 2.6 appears to show evidence of hour-to-hour variability in the log-transformed  $NO_2$  concentration. The concentration level increases suddenly at around 6 a.m. in every boxplot, as this is when the morning rush hour begins. The increase, however, is much lower for the Waulkmillglen Reservoir boxplot than for any of the others. This is most probably due to its distance away from any major roads and it being the only monitoring station in Glasgow City that was classified as ‘rural’.

### 2.1.3 DEFRA Modelled Grids

Figure 2.7(a) displays the  $NO_2$  pollution surface created by the 1 km by 1 km grids from the DEFRA-specified atmospheric dispersion model across the Greater Glasgow region. Figure 2.7(b) shows the pollution surface for the Glasgow City local authority. Figure 2.8 displays similar maps for  $PM_{10}$  concentrations with those for  $NO_X$  and  $PM_{2.5}$  being displayed in Appendix A. All maps are an annual average for the year 2011. The measurement units are micrograms per cubic metre ( $\mu\text{g}/\text{m}^3$ ).

Figure 2.7(a) shows that the highest concentrations of  $NO_2$  are present in the highly-urbanised areas of Glasgow City and its immediate surroundings as well as major roads such as motorways throughout the rest of the region. Figure 2.7(b) provides more detailed information regarding the spatial distribution of  $NO_2$  within Glasgow itself, with much higher concentrations being observed in the Central Business District and lower values being observed in peripheral areas. Similar inference can be made regarding the distribution of  $PM_{10}$  concentrations. However, as shown in Figures 2.7(b) and 2.8(b), there is a grid in the north-west of the city with unusually-high concentrations of  $NO_2$  and  $PM_{10}$ , respectively, being predicted. This could be attributed to local factors such as close proximity to a source of high air pollution.

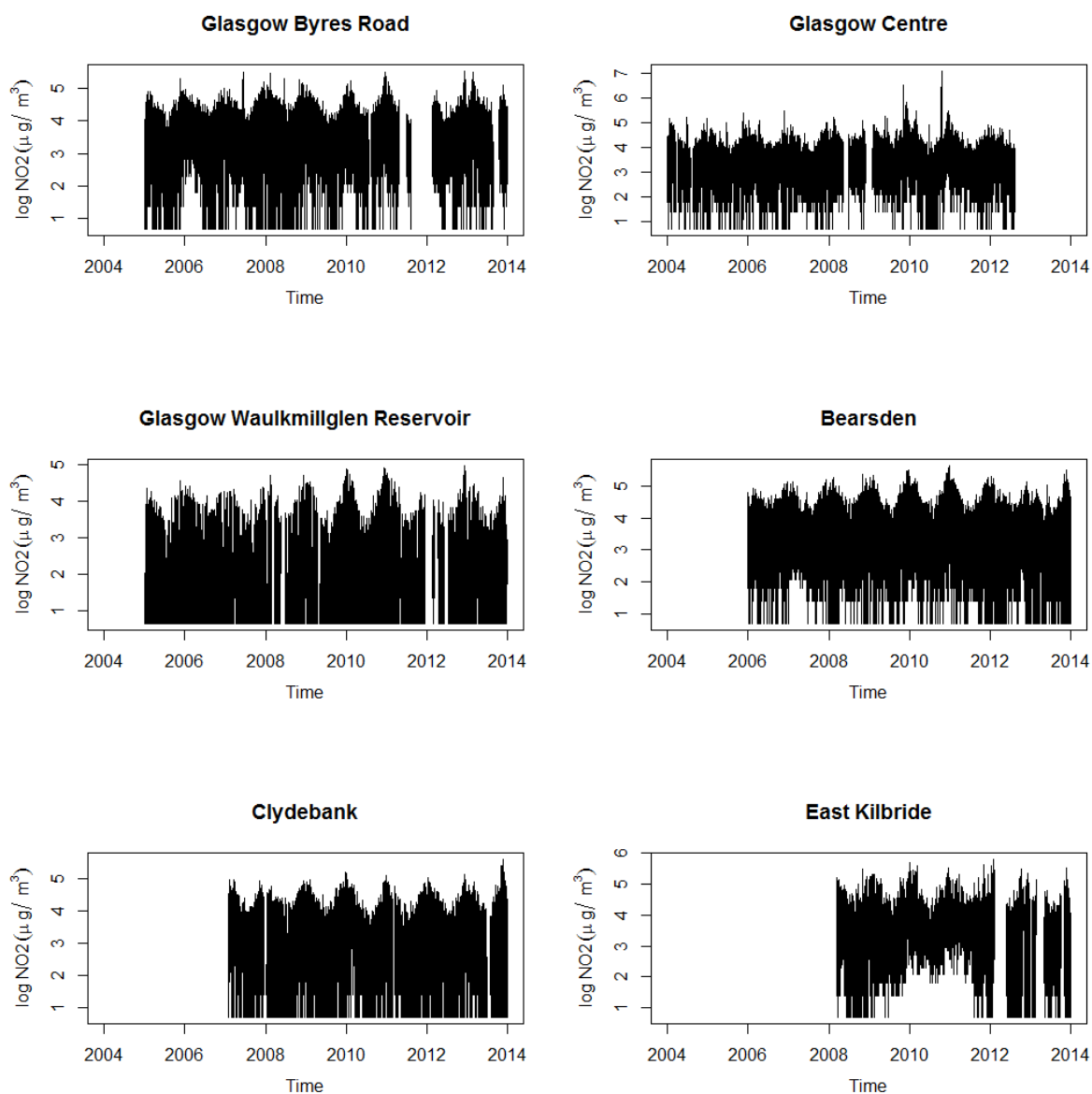


Figure 2.3: Time series plots of log-transformed  $NO_2$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

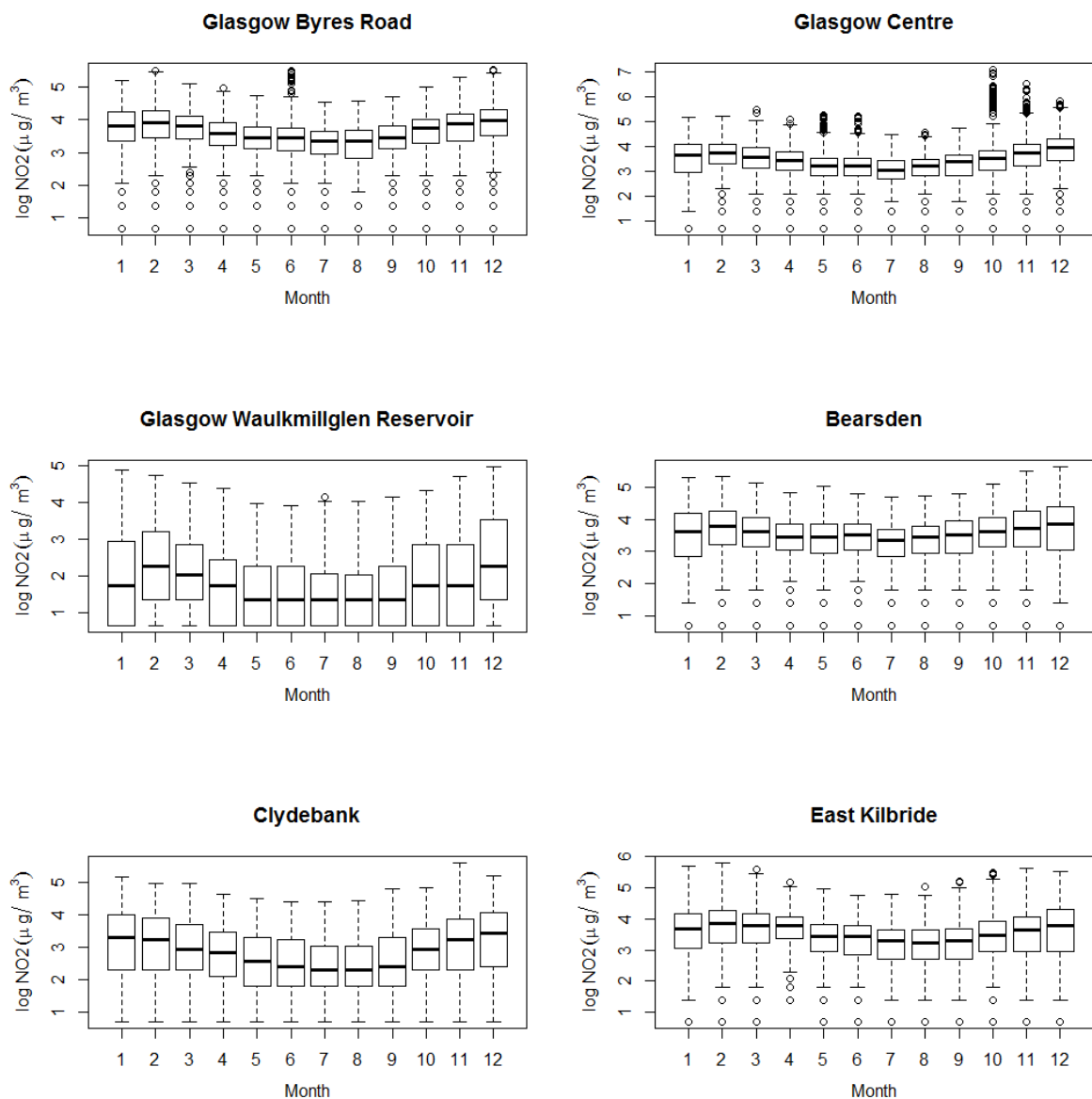


Figure 2.4: Boxplots of monthly log-transformed  $NO_2$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

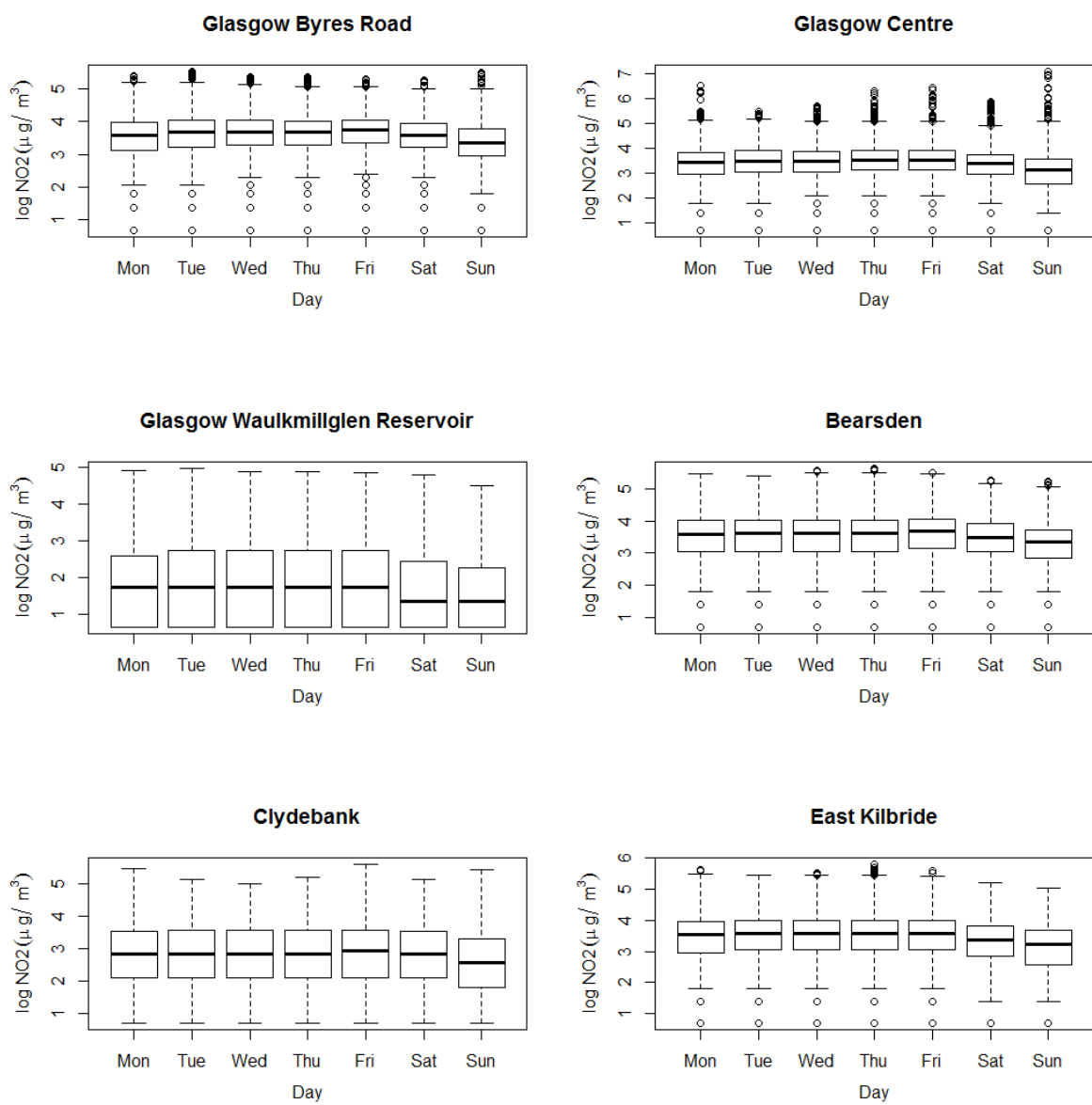


Figure 2.5: Boxplots of day-within-week log-transformed  $NO_2$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

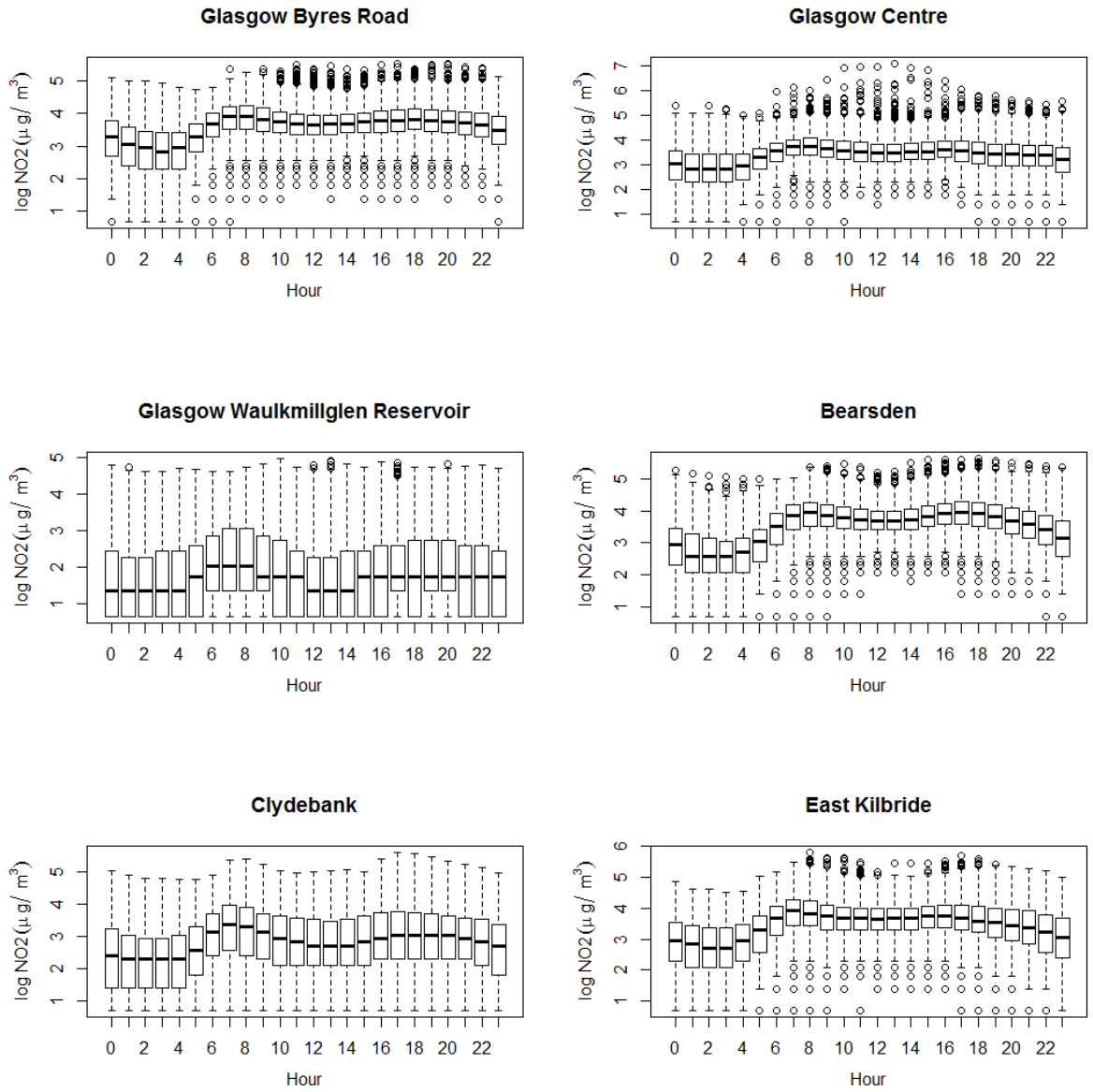


Figure 2.6: Boxplots of hourly log-transformed  $NO_2$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

Table 2.2 displays descriptive statistics for the  $NO_2$ ,  $NO_X$ ,  $PM_{10}$  and  $PM_{2.5}$  grids for the year 2011.

Table 2.2: Descriptive statistics for DEFRA modelled grids (2011 – units of measurement =  $\mu\text{g}/\text{m}^3$ )

Pollutant	Mean	St. Dev.	Min	Q1	Median	Q3	Max
$NO_2$	7.24	4.72	2.99	3.86	5.34	9.13	41.19
$NO_X$	9.68	6.98	3.80	4.91	6.83	12.08	76.88
$PM_{10}$	10.89	1.51	8.99	9.69	10.55	11.71	19.95
$PM_{2.5}$	7.31	0.86	6.25	6.69	6.99	7.65	12.34

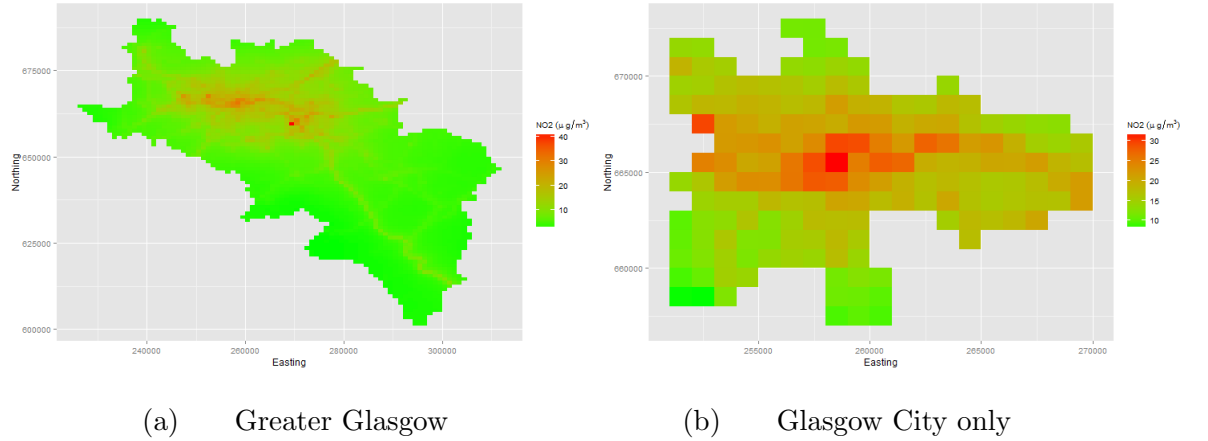


Figure 2.7: Maps with average 2011  $NO_2$  concentrations ( $\mu\text{g}/\text{m}^3$ ) for each 1 km by 1 km grid

#### 2.1.4 $NO_2$ Diffusion Tube Data

Local authorities throughout the U.K. also measure nitrogen dioxide levels in their ambient air by using diffusion tubes placed throughout their area. A diffusion tube is made of plastic and is approximately 7.5 cm long and 1 cm in diameter, with each end sealed by a cap (Doncaster Metropolitan Borough Council, 2014). Typically, one cap will be white and the other will be coloured. Inside the coloured cap there is a metal grid that has been soaked with triethanolamine or TEA, which absorbs nitrogen dioxide (Doncaster Metropolitan Borough Council, 2014).

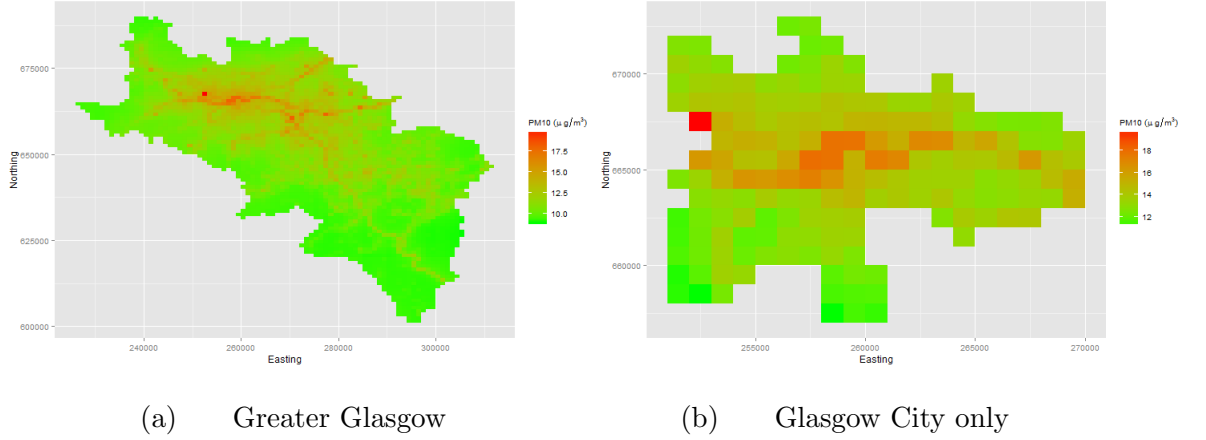


Figure 2.8: Maps with average 2011  $PM_{10}$  concentrations ( $\mu\text{g}/\text{m}^3$ ) for each 1 km by 1 km grid

Data are typically collected once per month, by removing the white cap and placing the diffusion tube near the area of interest, usually next to a busy road, using a plastic holder and a tie wrap (Doncaster Metropolitan Borough Council, 2014). The time during which the tube is open is known as the ‘exposure period’.

After the data collection process is completed, the white cap is replaced and the diffusion tube is then analysed in a laboratory to record the level of nitrogen dioxide that has been absorbed by the triethanolamine. The result is the average  $NO_2$  concentration in the air at that location for one month (Doncaster Metropolitan Borough Council, 2014).

Like the air monitoring data, the  $NO_2$  diffusion tube locations are divided into various categories, determined by their proximity to a major road and where exactly it is placed in the street, if applicable. The categories include ‘kerbside’, ‘roadside’, ‘urban centre’ and ‘urban background’, amongst others.

Table 2.3 shows descriptive statistics for the data. The units of measurement are micrograms per cubic metre ( $\mu\text{g}/\text{m}^3$ ).



Table 2.3: Descriptive statistics for  $NO_2$  diffusion tube data (units of measurement =  $\mu\text{g}/\text{m}^3$ )

n	Mean	St. Dev.	Min	Q1	Median	Q3	Max
311	32.83	12.81	5.90	24.00	31.00	39.18	99.60

Figure 2.9(a) shows the spatial map of the diffusion tube locations. Figure 2.9(b) displays the data points within the limits of Glasgow City only.

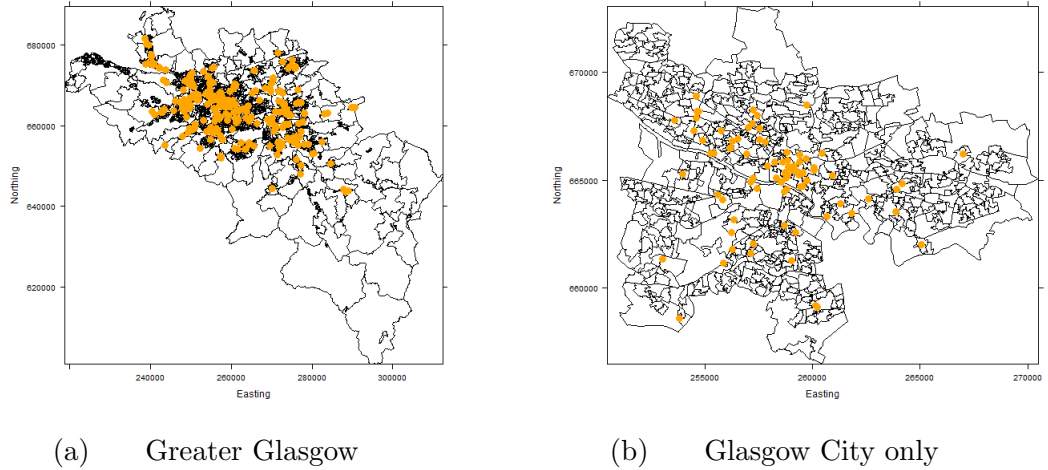


Figure 2.9: Map of  $NO_2$  diffusion tube locations in Greater Glasgow – contains Ordnance Survey data © Crown copyright and database right (2014)

Figure 2.10 shows that the median log-transformed  $NO_2$  concentration from the diffusion tubes does appear to change from year to year but any long-term increase or decline is not particularly visible.

### 2.1.5 Exploratory Spatial Analysis of All Air Quality Data

The three separate air quality data sets (air monitoring stations, modelled grids and diffusion tubes – the latter for  $NO_2$  only) were combined so that an exploratory spatial analysis could be undertaken, the results of which are displayed in Figure 2.11.

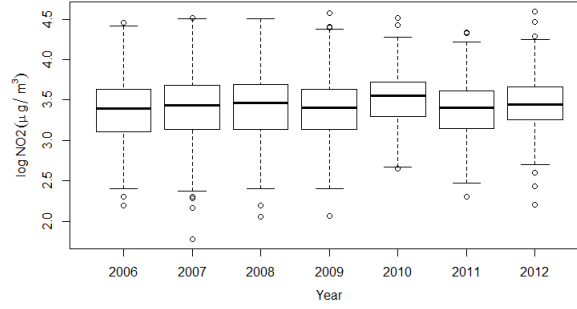


Figure 2.10: Boxplots of log-transformed  $NO_2$  concentrations ( $\mu\text{g}/\text{m}^3$ ) by year

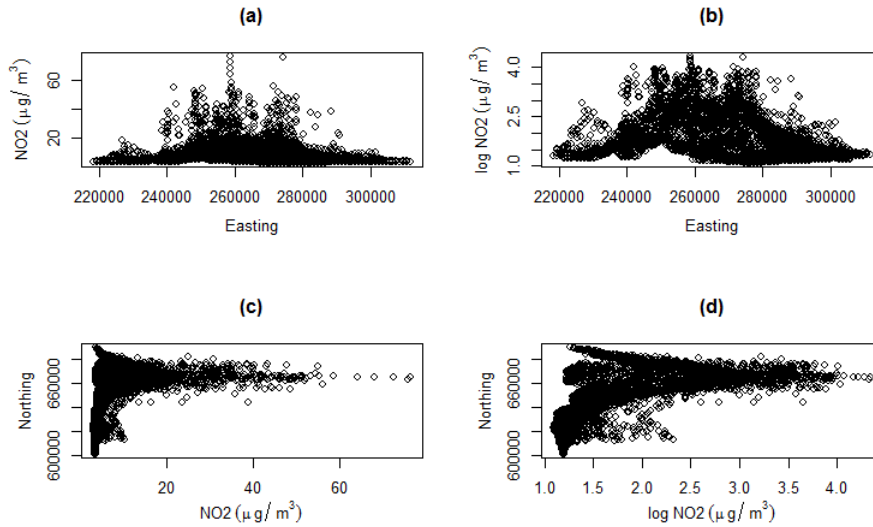


Figure 2.11: (a) scatterplot of easting versus  $NO_2$  concentration ( $\mu\text{g}/\text{m}^3$ ); (b) scatterplot of easting versus log-transformed  $NO_2$  concentration ( $\mu\text{g}/\text{m}^3$ ); (c) scatterplot of northing versus  $NO_2$  concentration ( $\mu\text{g}/\text{m}^3$ ); (d) scatterplot of northing versus log-transformed  $NO_2$  concentration ( $\mu\text{g}/\text{m}^3$ )

Figures 2.11(a) and 2.11(c) display  $NO_2$  concentration plotted against the Easting and Northing geographical coordinates, respectively. The  $NO_2$  concentrations were then log-transformed to allow for a clearer visual inference from the plots. Figure 2.11(b) shows that the log-transformed  $NO_2$  concentration appears to increase until the geographical midpoint of the study region, approximately, before returning to

initial levels. This may be evidence of a west-to-east pattern. Likewise, the log-transformed  $NO_2$  concentration plotted against Northing in Figure 2.11(d) shows an obvious increase in  $NO_2$  concentration as one moves in a southerly direction, again until reaching the centre of the study region.

Similar plots for  $NO_X$ ,  $PM_{10}$  and  $PM_{2.5}$  are displayed in Appendix A.

## 2.2 Water

The data used in this project were provided by the Scottish Environment Protection Agency (SEPA).

The data consisted of fifty-five determinands in a number of waterbodies and river catchments throughout the Greater Glasgow region. A map of these waterbodies is displayed in Figure 2.12(a) with a focus on Glasgow City in Figure 2.12(b). The determinands included chemical elements and compounds, biological data (macroinvertebrate levels) and general water characteristics such as temperature and pH. The data locations in Figure 2.12(a) are mostly within the River Clyde catchment but others are within either the catchments of the River Kelvin or the River Cart.

Each waterbody also received chemical, ecological and overall classifications dependent upon its scores for each of the recorded determinands. The overall water chemistry was classified as either pass/fail and the overall water ecology was classified as bad/poor/moderate/good/high, as mandated by the Water Framework Directive (European Union, 2000). The overall waterbody was classified according to a larger ordinal scale.

### 2.2.1 Water Determinands

The analysis focussed on six particular determinands that are particularly associated with water quality. The choices were made in consultation with the Scottish

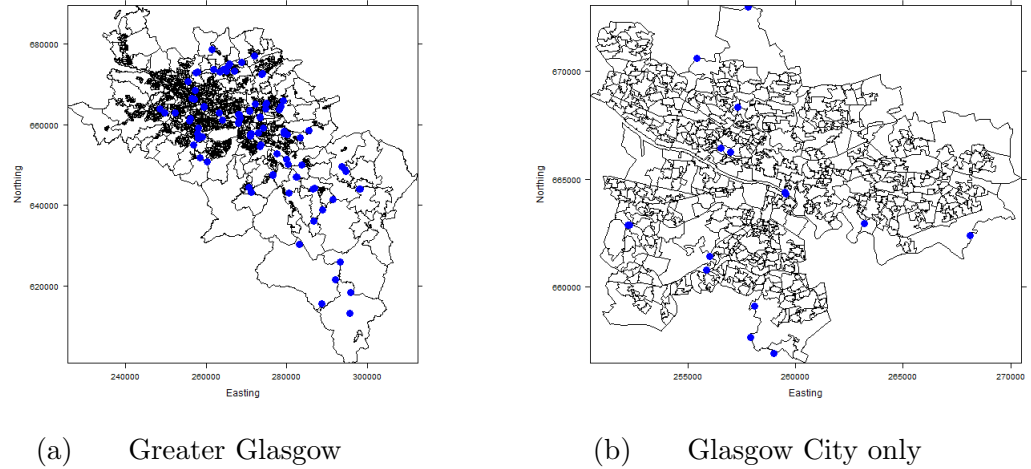


Figure 2.12: Map of data collection points – contains Ordnance Survey data

© Crown copyright and database right (2014)

Environment Protection Agency although no determinands that had been measured at less than forty spatial locations were included.

### **Ammonium ( $NH_4^+$ )**

The Scottish Environment Protection Agency (2015) describes ammonium as being both naturally-occurring (usually as a result of decaying natural matter) and present in water as a result of anthropogenic activities (such as fertiliser production and petrochemicals). It is colourless, corrosive and highly soluble in water. Although low concentrations are essential for many natural processes, high concentrations of ammonium are particularly harmful to waterbodies due to its toxicity with regards with aquatic wildlife and plants (Scottish Environment Protection Agency, 2015). Ammonium concentration was measured as milligrams per cubic litre (mg/L).

### **Copper ( $Cu$ )**

Copper is an insoluble reddish-brown metal (Health Protection Agency, 2003) and is found in both soils and waterbodies. It can be released into the environment from either natural sources such as volcanic eruptions and decaying vegetation or through anthropological activities such as mining and related industrial processes (Health

Protection Agency, 2003). Although a very small level of copper is present in all living organisms, excessive exposure is damaging to both human and environmental health (Health Protection Agency, 2003). Copper concentration was measured as grams per litre (g/L).

### **Dissolved Oxygen (*O*)**

Dissolved oxygen is critical to the survival of many aquatic lifeforms, particularly fish (Lewis and Evans, 2006). Typically, a lower level of dissolved oxygen is an indicator of a higher level of pollution in the water (Lewis and Evans, 2006). Oxygen enters the watercourse either through the atmospheric dissolution of the element at the surface or through the photosynthesis processes of nearby trees and plants (Lewis and Evans, 2006). The dissolved oxygen concentration was measured as % saturation.

### **pH**

pH is “a measure of the acidity” (Lewis and Evans, 2006) of a liquid and is ranked on a scale from 0 to 14. Solutions that have a pH value of less than 7 are designated as acidic and those above 7 are designated as basic or alkaline (Lewis and Evans, 2006). Solutions that have a pH value of 7 (i.e. pure water) are neutral. As the pH value decreases or increases from 7 in either direction, the more acidic or alkaline, respectively, it becomes. Water with a pH value of less than 3 is particularly dangerous to most forms of aquatic life (Lewis and Evans, 2006).

### **Soluble Reactive Phosphorus (*P*)**

Phosphorus is essential for aquatic plants and animals but is in short supply in most freshwater systems (United States Environmental Protection Agency, 2012b). However, modest increases in phosphorus levels can result in negative effects for rivers and streams such as accelerated plant growth, algae blooms, low dissolved oxygen (discussed above) and the death of fish, macroinvertebrates and other aquatic wildlife (United States Environmental Protection Agency, 2012b). In natural systems, elemental phosphorus is rare and mostly exists as part of an organic phosphate ( $PO_4$ )

molecule that is soluble (United States Environmental Protection Agency, 2012b). ‘Reactive’ phosphorus refers to what is being measured when testing for the presence of phosphorus in water (United States Environmental Protection Agency, 2012b). The phosphorus concentration was measured in grams per litre (g/L).

### **Macroinvertebrates**

An important component of measuring river water quality is biological monitoring. The United States Environmental Protection Agency (2012a) states that the three most common biological organisms present in rivers and streams are “fish, algae and macroinvertebrates.” Macroinvertebrates are organisms such as insects, molluscs and worms that are visible to the naked eye (United States Environmental Protection Agency, 2012a) and are routinely used to monitor the ecological quality of rivers in Scotland (Scottish Natural Heritage, 2009). Most of these organisms are found attached to underwater rocks, logs and plants (United States Environmental Protection Agency, 2012a).

Macroinvertebrates are usually measured by the number of taxa or the average number per taxon, where taxa are the families or species within the entire community of macroinvertebrates in the waterbody (Scottish Natural Heritage, 2009). A greater number of macroinvertebrates in a recorded taxon typically indicates a higher level of water quality (Scottish Natural Heritage, 2009).

The United States Environmental Protection Agency (2012a) identifies reasons why macroinvertebrates are good indicators of river water quality:

- They respond to biological, chemical and physical conditions and changes in the stream
- They are susceptible to pollution events
- They may show impacts from habitat loss not detected by traditional water quality assessment methods

- They are a fundamental part of a river's food chain
- They are relatively easy to sample and identify

In the data set used in this project, macroinvertebrate activity was measured by calculating the average score per taxon (often abbreviated to ASPT).

## 2.2.2 Exploratory Analysis

Table 2.4 displays descriptive statistics for each water determinand (excluding macroinvertebrates).

Table 2.4: Descriptive statistics for water chemistry determinands

<b>Determinand (units)</b>	<b>n</b>	<b>Mean</b>	<b>St. Dev.</b>	<b>Min</b>	<b>Q1</b>	<b>Median</b>	<b>Q3</b>	<b>Max</b>
Ammonium (mg/L)	57	0.19	0.51	0.003	0.02	0.06	0.16	8.61
Copper (g/L)	57	149.68	114.08	6.32	81.10	119.00	181.00	2229.00
Dissolved Oxygen (%)	57	94.15	9.38	10.50	90.40	95.50	99.60	136.00
pH	57	7.76	0.33	6.12	7.55	7.78	7.99	8.77
Phosphorus (g/L)	57	0.08	0.11	0.002	0.02	0.05	0.09	1.31

A detailed exploratory analysis is only given here for the soluble reactive phosphorus determinand as well as the macroinvertebrates. Exploratory plots for the remaining four determinands are provided in Appendix A of this thesis.

### Soluble Reactive Phosphorus

Although there are fewer data points than the corresponding exploratory analysis for air quality, Figures 2.13(a) and 2.13(c) still show that a pattern in phosphorus concentration may be present. A log-transformation (shown in Figures 2.13(b) and 2.13(d)) highlights this, with higher concentrations being observed in the centre of the study region, at least for the Northing variable. A pattern is less evident for the Easting variable. Figure 2.14 provides more detail on the distribution of the data.

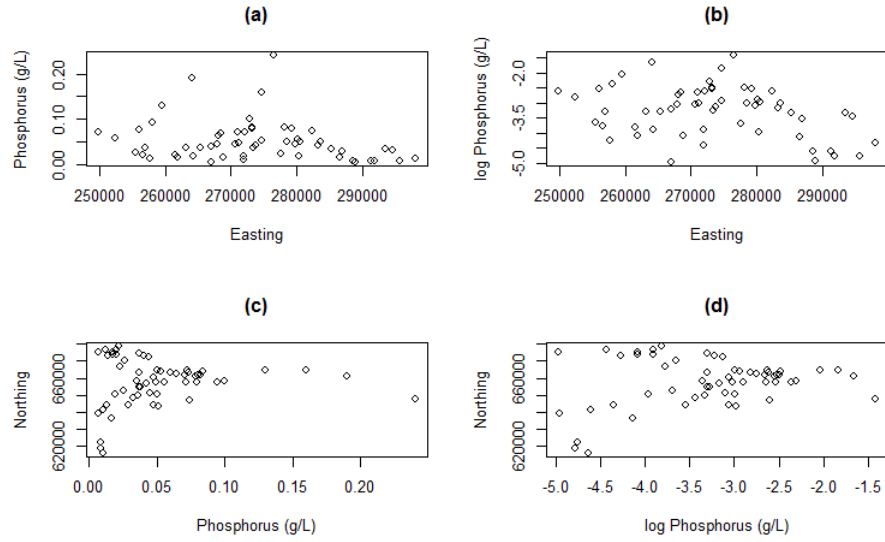


Figure 2.13: (a) plot of Easting versus phosphorus concentration (g/L); (b) plot of Easting versus log-transformed phosphorus concentration (g/L); (c) plot of phosphorus concentration (g/L) versus Northing; (d) plot of log-transformed phosphorus concentration (g/L) versus Northing

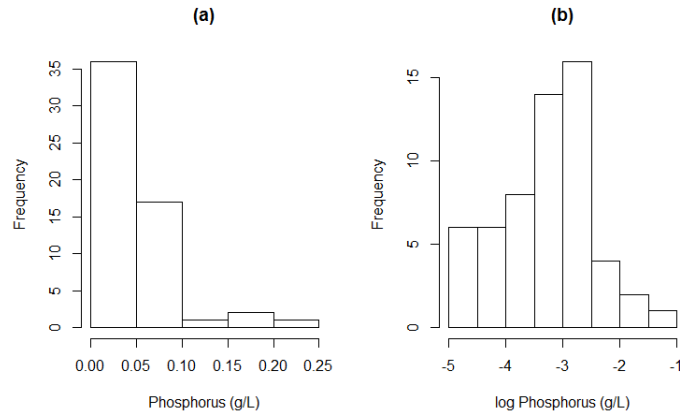


Figure 2.14: (a) boxplots of annual log-transformed phosphorus concentration (g/L); (b) boxplots of monthly log-transformed phosphorus concentration (g/L)

Figure 2.14(a) shows that the data are highly skewed to the right and that a transformation might be justified. Figure 2.14(b) displays the corresponding log-transformation of phosphorus concentration. Although there remains a heavy lower



tail, the log-transformed data are much more bell-shaped than in Figure 2.14(a). A temporal exploratory analysis of the data is shown in Figure 2.15.

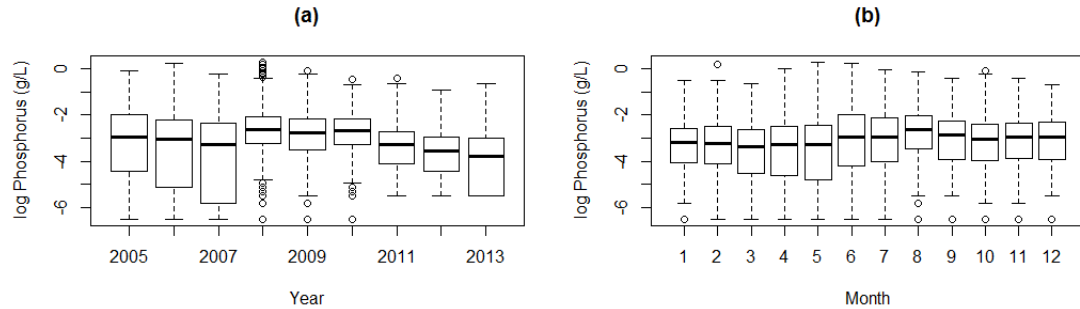


Figure 2.15: (a) boxplots of annual log-transformed phosphorus concentration (g/L); (b) boxplots of monthly log-transformed phosphorus concentration (g/L)

As shown in Figure 2.15(a), the log-transformed phosphorus concentration does not appear to exhibit much of a temporal trend. Figure 2.15(b) does not show much evidence of a seasonal pattern either albeit with marginally higher concentrations typically being observed in July and August.

### Macroinvertebrates – Average Score per Taxon

The macroinvertebrate data were classified on an ordinal scale as opposed to a continuous scale with the average score per taxon being given a score of ‘bad’, ‘poor’, ‘moderate’, ‘good’ or ‘high’. Figure 2.16 shows the proportions of the five classification levels according to the average scores of macroinvertebrate numbers per taxon. This is shown annually to allow for temporal comparisons.

The majority of the taxa received scores of ‘moderate’, ‘good’ or ‘high’, meaning that the average macroinvertebrate numbers are at least adequate, if not better in many locations. There has been year-on-year variations in these proportions with 2012 being a particularly good year for macroinvertebrate numbers, with almost no locations receiving the worst classification. There was however, a return to previous

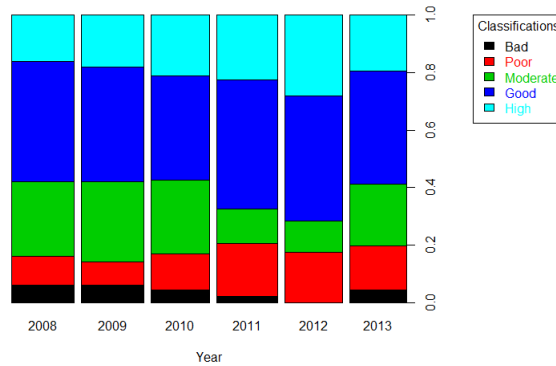


Figure 2.16: Annual proportions of macroinvertebrate classifications (using the average score per taxon)

numbers in 2013, perhaps showing that there is little to no long-term trend in the average number of macroinvertebrates per taxon.

Figure 2.17 is a map of the macroinvertebrate sampling locations and the corresponding classification allocated to each data point.

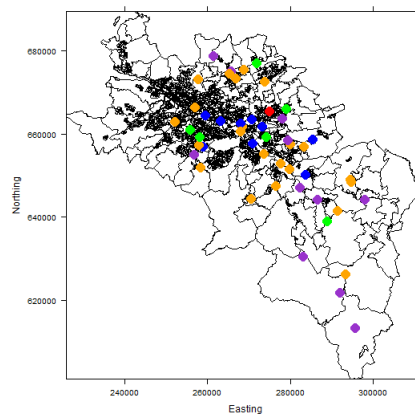


Figure 2.17: Map of macroinvertebrate sampling locations and classifications (● = bad, ● = poor, ● = moderate, ● = good, ● = high) – contains Ordnance Survey data © Crown copyright and database right (2014)

Table 2.5 provides a summary of river catchment information as well as the number of cases for each of the five categories.

Table 2.5: Status and catchment information for macroinvertebrate data

<b>Status</b>	<b>River Clyde</b>	<b>River Kelvin</b>	<b>River Cart</b>	<b>Total</b>
Bad	1	0	0	1
Poor	8	0	1	9
Moderate	3	1	2	6
Good	13	6	3	22
High	8	2	1	11
<b>Total</b>	33	9	7	49

## 2.3 Land

The data used in this project were kindly provided under licence from the British Geological Survey, as part of the Geochemical Baseline Survey of the Environment (G-BASE) Glasgow Soil data set BGS, © NERC (Fordyce et al., 2012), which sought to provide an overview of the urban soil geochemistry of Glasgow and the immediate suburban areas.

The data set consisted of topsoil concentrations of five metals (arsenic, chromium, lead, nickel and selenium) at 1,622 study sites across the Greater Glasgow conurbation. All samples were collected between 2001 and 2002 with 1,381 of the samples being classified as ‘urban’ at a density of 1 per 0.25 km<sup>2</sup> and 241 as ‘peri-urban’ at a density of 1 per 2 km<sup>2</sup>. All five metals were sampled at each site. The map of the 1,622 study sites is displayed in Figure 2.18(a) with those within the limits of Glasgow City being displayed in Figure 2.18(b).

The data that were provided were subject to a lower limit of detection determined by the British Geological Survey. These limits are displayed in Table 2.6 alongside the corresponding element.

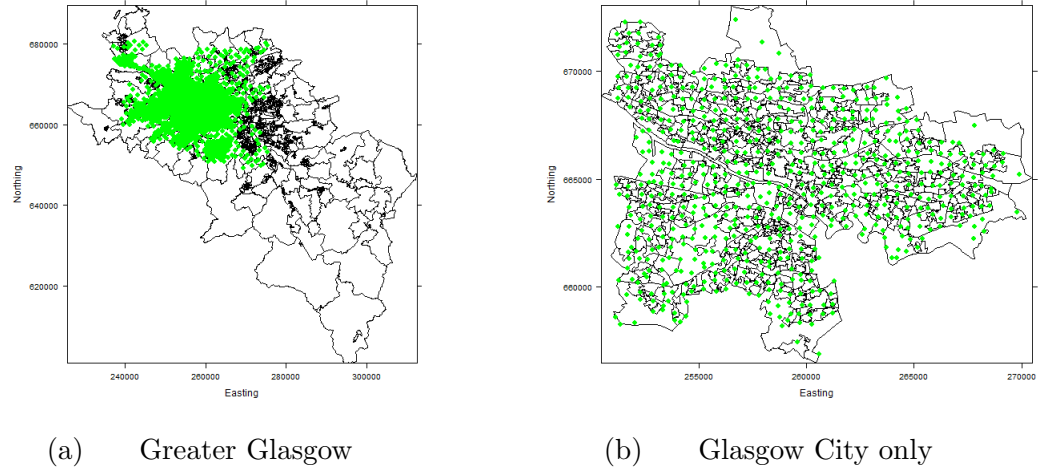


Figure 2.18: Map of sites for G-BASE project (Fordyce et al., 2012) – contains Ordnance Survey data © Crown copyright and database right (2014)

Table 2.6: Soil determinands and limits of detection (British Geological Survey, 2014) – units of measurement = mg/kg

Element	Name	Limit of Detection
As	Arsenic	0.9
Cr	Chromium	1.3
Ni	Nickel	0.6
Pb	Lead	0.5
Se	Selenium	0.2

The Scottish Environment Protection Agency (2014) defines these five metals as ‘potentially toxic elements’. They often enter topsoils through natural environmental processes and are beneficial to the environment in very small quantities (Scottish Environment Protection Agency, 2014). However, in high concentrations they reduce crop yield and have adverse health effects in both humans and animals. There is also the possibility of unsafe concentrations of these and other metals entering the human food chain if present in agricultural land.

### 2.3.1 Soil Pollutants

Below is a discussion of all five metals in the G-BASE project:

**Arsenic (*As*)**

Arsenic occurs in trace quantities in all rock, soil, water and air (World Health Organisation, 2001). Mining, smelting of nonferrous metals and burning of fossil fuels are the main anthropogenic contributors to arsenic-related contamination of air, water and soil (Berkowitz et al., 2014). Arsenic concentrations were measured in milligrams per kilogram (mg/kg).

**Chromium (*Cr*)**

Airborne chromium compounds are present mostly as fine dust particles, which settle over water and often strongly attach themselves to soil (Berkowitz et al., 2014). It is widely regarded as carcinogenic (Berkowitz et al., 2014). Chromium concentrations were measured in milligrams per kilogram (mg/kg).

**Nickel (*Ni*)**

Nickel is “a silver-white, lustrous, hard, malleable, ductile and ferromagnetic metal that is resistant to corrosion and a reasonable conductor of electricity and heat” (Berkowitz et al., 2014), comprising approximately 0.008% of the Earth’s crust (Berkowitz et al., 2014). Nickel is emitted to the environment from both natural and anthropogenic sources such as industrial processes and waste disposal (Berkowitz et al., 2014). Nickel concentrations were measured in milligrams per kilogram (mg/kg).

**Lead (*Pb*)**

Lead is “a blue-white lustrous metal that is very soft, malleable, ductile and resistant to corrosion but tarnishes upon exposure to air” (Berkowitz et al., 2014). A wide distribution of lead in soils has been reported with an average lead content of 10 mg/kg in topsoil (United States Environmental Protection Agency, 1992). Lead poisoning is an environmental and public health hazard of global proportions and can result of one high-level exposure or multiple high or low-level exposures (Berkowitz et al., 2014). Lead concentrations were measured in milligrams per kilogram (mg/kg).

## Selenium (*Se*)

Selenium is widely, though unevenly, distributed over the surface of the Earth with an average 50 to 200 mg/kg (milligrams per kilogram) concentration in soils (Reilly, 2006). Used in lasers, plastics, ceramics, paints and in agriculture and horticulture. It is an essential component of the human diet, though only in minute amounts. If these amounts are exceeded, it can have severe adverse effects on human health (Reilly, 2006). Selenium concentrations were measured in milligrams per kilogram (mg/kg).

### 2.3.2 Exploratory Analysis

Table 2.7 displays descriptive statistics in relation to each of the five metals in the G-BASE data set.

Table 2.7: Descriptive statistics of metals in G-BASE data set (units of measurement = mg/kg)

<b>Element</b>	<b>Mean</b>	<b>Sd. Dev.</b>	<b>Min</b>	<b>Q1</b>	<b>Median</b>	<b>Q3</b>	<b>Max</b>
Arsenic	10.79	10.56	1.10	7.30	9.10	11.40	282.80
Chromium	121.56	130.31	28.00	91.00	107.00	126.00	4286.00
Lead	167.91	210.49	13.40	77.63	118.30	187.50	5001.00
Nickel	51.56	43.75	2.30	35.00	45.70	59.08	1038.10
Selenium	1.03	0.68	0.10	0.70	0.90	1.20	14.50

The results of an exploratory analysis for one metal (lead) are displayed below. Such plots for the remaining four metals are provided in Appendix A.

The plots of lead concentration against Easting and Northing in Figures 2.19(a) and 2.19(c), respectively, offer little inferential value. Figure 2.19(b), the log-transformed equivalent of Figure 2.19(a), shows that lead concentration does not vary much in a west-to-east direction. However, a pattern may exist in Figure 2.19(d), the log-transformed equivalent of Figure 2.19(c), where the lead concentration appears to increase as the Northing coordinate approaches the centre of the study region, before declining again towards the far south of Greater Glasgow.

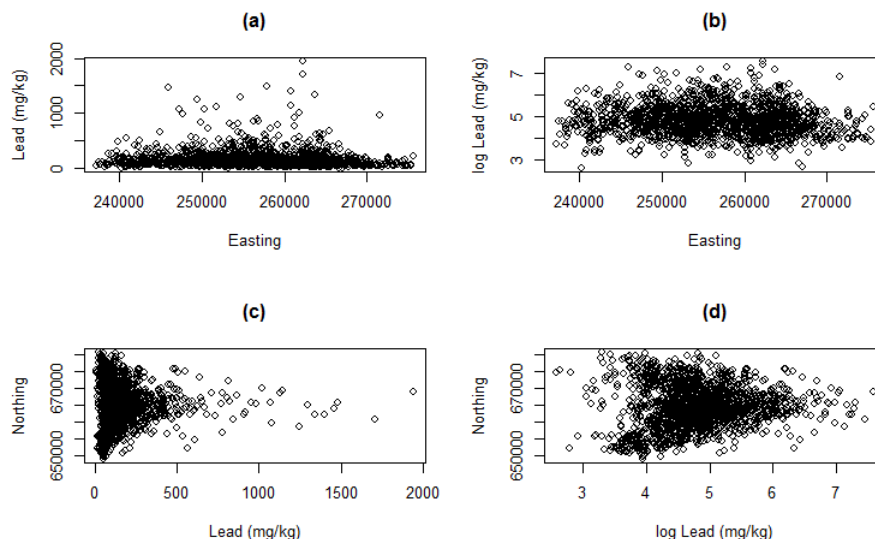


Figure 2.19: (a) scatterplot of easting versus lead concentration (mg/kg); (b) scatterplot of easting versus log-transformed lead concentration (mg/kg); (c) scatterplot of northing versus lead concentration (mg/kg); (d) scatterplot of northing versus log-transformed lead concentration (mg/kg)

Figure 2.20 shows evidence of why a log-transformed response variable would provide better inference in the modelling procedures to be subsequently undertaken. The log-transformed metal concentration shown in Figure 2.20(b) appears to be more symmetric and bell-shaped, especially when compared to the corresponding histograms for the air pollutants.

## 2.4 Temporal Frequency of Data

Table 2.8 summarises the temporal frequency of the available data. It should be noted that this is a general summary and almost all available data are subject to periods of missing data. For the water determinands, it is also common in places for two measurements to be taken monthly (one near the beginning of the month and one near the end) and then no measurement taken the following month at all. For simplicity and regularity, the second measurement nearer the end of the month

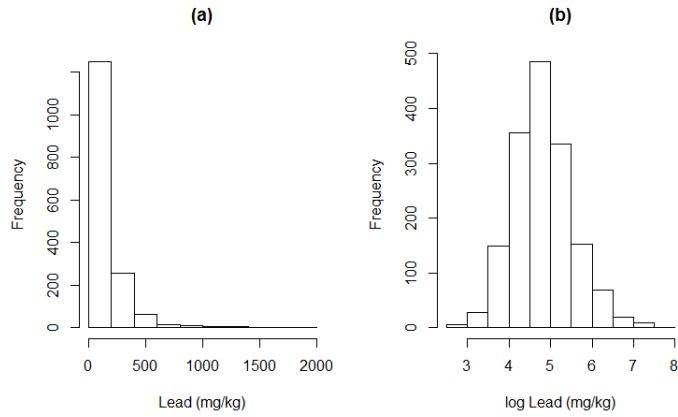


Figure 2.20: (a) histogram of lead concentration (mg/kg) values; (b) histogram of log-transformed lead concentration (mg/kg) values

could be treated as a measurement for the following month where no sample was taken.

Table 2.8: Temporal frequency of data

<b>Data</b>	<b>Annual</b>	<b>Monthly</b>	<b>Daily</b>	<b>Hourly</b>
Air Monitoring Stations	✓	✓	✓	✓
DEFRA Modelled Grids	✓			
$NO_2$ Diffusion Tubes	✓	✓		
Water	✓	✓		
Soil	✓			

All statistical processes using the data in the following chapters will involve an annual average only as this is the minimum temporal resolution for which air, water and soil data are available. For the air and water quality data, annual averages should also eliminate the issue of potential outliers, the effect of which will be greatly reduced by using an annual mean.

## 2.5 Chapter Summary

The air quality data used in this project were downloaded from public sources and included data from monitoring stations and  $NO_2$  diffusion tubes in Greater Glasgow



as well as modelled grids from an atmospheric dispersion model developed by the Department for Environment, Food and Rural Affairs. When all three sources of data were considered together, data were available for four air pollutants: nitrogen dioxide ( $NO_2$ ), nitric oxide ( $NO_X$ ) and particulate matter ( $PM_{10}$  and  $PM_{2.5}$ ).

For the monitoring stations, an exploratory analysis was undertaken for the year 2011, which consisted of basic summary statistics (such as the mean, standard deviation, median etc.), time series plots to identify if a long-term trend was present and boxplots of the data plotted against month, day of the week and hour, respectively, to check if a seasonal pattern existed. The distribution of the data were also assessed by the use of a histogram.

An exploratory analysis of the  $NO_2$  diffusion tubes comprised the use of summary statistics and boxplots for annual concentrations. Spatial concentration maps were constructed for the DEFRA modelled grids as well as summary statistics for these data. All of the air quality data were plotted against geographic coordinates (east-ing and northing, respectively) to informally assess whether a spatial pattern was present.

The water quality data were provided by the Scottish Environment Protection Agency. Five continuous, chemical determinands (ammonium, copper, dissolved oxygen, soluble reactive phosphorus and pH) and one categorical, biological determinand (macroinvertebrates – measured as the average score per taxon) were selected in consultation with SEPA. For the former, summary statistics and plots were constructed in a manner similar to that for the air pollutants. For the macroinvertebrate data, the changes in classifications over time were assessed, as was the number of classifications within each of the three river catchments.

The soil quality data were provided by the British Geological Survey and consisted of five pollutants: arsenic, chromium, lead, nickel and selenium. An exploratory

analysis of these data consisted of summary statistics and an informal assessment of the spatial patterns present. The distributions of the data were also assessed.

All data across the three domains (excluding macroinvertebrates) were log-transformed in order to satisfy the normality assumptions that would be made in later chapters. Finally, since the soil data and DEFRA grids were only available at an annual level, any subsequent analyses would only be able to be undertaken at the same temporal scale.

# Chapter 3

## Geostatistical Modelling of Air, Water and Soil Data

This chapter involves the modelling of the data introduced in Chapter 2 as a step towards the construction of a composite index for environmental quality in Greater Glasgow. This is achieved by finding a suitable model that describes the spatial patterns exhibited by the data and then drawing predictions from these smooth spatial surfaces in order to allow for ‘data’ to be ‘observed’ in every data zone in the study region.

### 3.1 Methods

#### 3.1.1 Flexible Regression

A normal linear regression model where each covariate is constrained to be a linear predictor is often insufficient for modelling environmental and other data sets and alternative models that relax the rather strict assumptions of standard linear regression need to be considered (Green and Silverman, 1994). Data that show evidence of curvature in the response variable as opposed to linearity can be modelled in a variety of ways. A key method is the generalised additive model (Hastie and Tibshirani, 1990) which can model linear or nonparametric relationships between a

response variable and its predictors and also incorporates the theory of generalised linear models by allowing the response to be modelled using a distribution which is part of the exponential family. When the response is assumed to follow a Gaussian distribution, the generalised additive model reduces to an additive model.

In the normal linear model, all covariates are assumed to have a linear relationship with the response variable as well as assuming that the error term follows a Gaussian distribution with mean zero and constant variance (homoscedasticity), i.e.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (3.1)$$

where  $\mathbf{Y}$  is an  $(n \times 1)$  vector of response variables,  $\mathbf{X}$  represents an  $(n \times p)$  matrix of model coefficients,  $\boldsymbol{\beta}$  is an  $(p \times 1)$  vector of parameters and  $\boldsymbol{\epsilon}$  is an  $(n \times 1)$  vector of error terms ( $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ ).

Where the response variable shows evidence of curvature, one way to account for this is to replace the linear predictors with nonparametric functions, such as splines (Ruppert et al., 2003). Not all covariates will necessarily require this alteration and some can be left as linear predictors if they show evidence of a linear relationship with the response variable. This nonparametric or semiparametric (where linear predictors remain) model is an additive model (Hastie and Tibshirani, 1990), where the error terms adhere to the same distributional assumptions as that for a normal linear model.

An additive model can be expressed mathematically as:

$$y_i = \beta_0 + \sum_{t=1}^m f_t(x_{ti}) + \epsilon_i \quad (3.2)$$

where  $t = 1, \dots, m$  explanatory variables and  $\epsilon_i \sim N(0, \sigma^2)$ .  $\beta_0$  is the intercept

term,  $f_1(x_1), \dots, f_m(x_m)$  are smooth functions of the covariates and the identifiability constraint  $\sum_{i=1}^n f_t(x_{ti}) = 0$  (Green and Silverman, 1994).

An additive model can be fitted using the backfitting algorithm, with components such as smoothing splines, or through a penalised regression splines approach.

Hastie and Tibshirani (1990) define the smoothing spline formulation, for one covariate, as:

$$S(f) = \sum_{i=1}^n (Y_i - f(t_i))^2 + \lambda \int_a^b (f''(x))^2 dx \quad (3.3)$$

The penalised least squares estimator  $\hat{f}$  is “defined to be the minimiser of the function  $S(f)$  over the class of all twice-differentiable functions  $f$ ” (Green and Silverman, 1994). The roughness penalty  $\lambda \int_a^b (f''(x))^2 dx$  is included to account for the curvature in the data and, for an appropriate value of the smoothing parameter  $\lambda$ , the minimisation of  $S(f)$  should be able to “compromise between smoothness and goodness-of-fit” (Green and Silverman, 1994).

Higher values for  $\lambda$  will result in a smoother fit for  $f$ . Hastie and Tibshirani (1990) observe that as  $\lambda \rightarrow \infty$ , the penalty term will dominate the expression leading to ever-increasing smoothness. This will ultimately result in the least squares solution. However, as  $\lambda \rightarrow 0$ , the penalty term diminishes and the model will interpolate the data. An appropriate choice can be found either through manual testing of various values of  $\lambda$  or through automatic methods such as cross-validation, generalised cross-validation or Akaike’s Information Criterion (AIC), amongst others.

Cross-validation is a very common method for finding an optimal value for  $\lambda$ . Hastie and Tibshirani (1990) describe its formulation. A cross-validation process moves through all data points, leaving out the  $i^{th}$  data point and then fitting the model using the remaining  $n - 1$  points. This is expressed mathematically (Hastie and

Tibshirani, 1990) as:

$$\text{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_{\lambda}^{-i}(x_i))^2 \quad (3.4)$$

where  $\hat{f}_{\lambda}^{-i}$  is the fit of the model at data point  $x_i$  with the  $i^{\text{th}}$  point being omitted.

However, Faraway (2006) notes that cross-validation is “computationally expensive” and a related method, generalised cross-validation, is often used as an alternative. The **mgcv** (Wood, 2006) package in R (R Core Team, 2014), for instance, calculates optimal smoothing parameters using generalised cross-validation by default.

Generalised cross-validation, for nonparametric regression (effectively an additive model with only one covariate), is defined by Ruppert et al. (2003) as follows:

$$\text{GCV}(\lambda) = \frac{\text{RSS}(\lambda)}{(1 - n^{-1}\text{tr}(S_{\lambda}))^2} \quad (3.5)$$

where  $n$  is the number of data points, RSS is the residual sum of squares, ‘tr’ is the trace and  $S_{\lambda}$  is the smoothing matrix. The generalised cross-validation formula for an additive model, which contains more than one explanatory variable, is:

$$\text{GCV} = \frac{n\text{RSS}}{(n - \text{tr}(P))^2} \quad (3.6)$$

where  $n$  is the number of data points, ‘tr’ is the trace,  $P$  is the projection matrix (representing the fit over all explanatory variables) and RSS is the residual sum of squares.

As aforementioned, another way to fit an additive model is to use regression splines (or penalised regression splines), which differ in that the number of knots is used to control the level of smoothing as opposed to the observed  $x$  values themselves

(Faraway, 2006). This method is used by the **mgcv** package (Wood, 2006) for additive models in R (R Core Team, 2014), which was primarily used for statistical analyses in this thesis.

Regression splines are comprised of basis functions. Ruppert et al. (2003) describe the adaptation of Equation 3.3 for smoothing splines to the penalised regression splines method.

$$S(f) = ||\mathbf{y} - X\boldsymbol{\beta}||^2 + \lambda^2 \boldsymbol{\beta}^T D \boldsymbol{\beta} \quad (3.7)$$

where  $\lambda \geq 0$  and  $\lambda^2 \boldsymbol{\beta}^T D \boldsymbol{\beta}$  is the roughness penalty.  $D$  is a matrix of basis functions. The level of smoothing in a penalised regression spline is determined by the basis dimension and the smoothing parameter  $\lambda$ . Typically, a large number of basis functions are used and  $\lambda$  controls the ‘wiggleness’ of the fit.

The  $k$ -index, featured in the **mgcv** package, can be used to select an appropriate level of smoothing. The  $k$ -index is an estimate of the residual variance in the data after the fitting of an additive model (Wood, 2006) with values below 1 being evidence of there being a significant pattern remaining in the data. Wood (2006) recommends that the  $k$ -index should be at least 1 and this can be achieved by increasing the number of knots used to fit the model.

It is also possible to create flexible regression models that contain smooths of several variables, known as isotropic smooths (Wood, 2006). In the **mgcv** package (Wood, 2006) in R (R Core Team, 2014), these models are fitted using thin-plate regression splines, which are “invariant to rotation of covariate axes” (Wood, 2006). One drawback is that they are not invariant to covariate rescaling (Wood, 2006). Wood (2006) recommends that such isotropic smooths are particularly appropriate where all components of the smooth are measured in the same units (i.e. Easting and Northing or longitude and latitude).

## Parameter Estimation

Smoothing splines for additive model components are typically fitted using the back-fitting algorithm, described by Faraway (2006).

Firstly, an initial value is chosen by the intercept term  $\beta_0$  and all  $f_j(x)$ . For the intercept, a typical choice is the sample mean,  $\bar{y}$  and for the smooth terms, typically the least squares estimates for  $j = 1, \dots, p$ .

Then, for  $j = 1, \dots, p, 1, \dots, p, \dots$ ,

$$f_j = S(x_j, y - \beta_0 - \sum_{i \neq j} f_i(X_i)) \quad (3.8)$$

where  $S(x, y)$  is the smooth of the data  $(x, y)$  and the algorithm is iterated until convergence. The term  $y - \beta_0 - \sum_{i \neq j} f_i(X_i)$  is a partial residual. Many computational smoothing splines algorithms (such as the **gam** package (Hastie, 2013) in R (R Core Team, 2014)) require manual specification of  $S$ , which could be a nonparametric function such as splines or loess (locally-weighted scatterplot smoother) or even a parametric linear or polynomial function (Faraway, 2006).

For regression splines, Wood (2006) observes that to estimate a function,  $f$ , “requires that  $f$  be represented such that it is a linear model.” This is achieved by selecting an appropriate basis, which defines “the space of functions of which  $f$  is an element” (Wood, 2006). Wood (2006) states that if  $b_j(x)$  is the  $j^{th}$  basis function, then

$$f(x) = \sum_{j=1}^q b_j(x) \beta_j \quad (3.9)$$

for some values of unknown parameters  $\beta_j$  and a total of  $q$  basis functions.

When fitting a spatial surface (termed an isotropic or bivariate smooth) in R (R



Core Team, 2014), the **mgcv** package (Wood, 2006) fits an additive model using penalised regression splines with an extension of the model in Equation 3.7.

In spatial contexts, particularly those with small sample sizes, it is common for restricted maximum likelihood or REML (Patterson and Thompson, 1971) to be used for parameter estimation as opposed to maximum likelihood. This approach will usually lead to less bias (Diggle and Ribeiro, 2007), especially for small sample sizes. Despite this, Diggle and Ribeiro (2007) also observe that REML can be more sensitive to the choice of a mean model than maximum likelihood.

### **Model Diagnostics and Residuals**

Ruppert et al. (2003) state that diagnostics refer to a “large collection of techniques used to check the quality of the data and the adequacy of a regression model.” Many of the traditional diagnostics procedures used for linear models such as consideration of fitted values and residuals are also useful for flexible regression models.

The  $i^{th}$  fitted value is the estimate of  $E(y_i)$  from the fitted model, i.e.  $\hat{y}_i = x_i^T \hat{\beta}$  (Ruppert et al., 2003).

The  $i^{th}$  residual is defined as  $e_i = y_i - \hat{y}_i$ , i.e. the difference between the observed  $i^{th}$  measurement and the corresponding  $i^{th}$  prediction from the fitted model (Ruppert et al., 2003). If the model is a good description of the data in question, then the residuals should have mean zero and constant variance. This is a consequence of the systematic component of the model effectively capturing the mean and variability of the data, leaving only random noise.

These diagnostics are assessed through a variety of plots, whereupon the distributional assumptions of the model (particularly its error component) can be checked.

### 3.1.2 Geostatistical Modelling

According to Diggle and Ribeiro (2007), a geostatistical process involves a number of  $Y_i$  ( $i = 1, \dots, n$ ) observed at corresponding locations  $x_i$  within some spatial region of interest. Also, each observed  $Y_i$  is a measurement taken at location  $x_i$  or a value relating to an underlying stochastic process  $T(x)$  at location  $x_i$ .

Sherman (2011) states that geostatistical modelling has two principal aims, namely to predict the spatial process at unobserved locations based on the data from observed locations and to understand how the process varies between “neighbouring locations.”

A geostatistical model for normally-distributed data can be defined as:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \quad (3.10)$$

where  $\mathbf{Y}$  is an  $(n \times 1)$  vector of response variables,  $\boldsymbol{\mu}$  is an  $(n \times p)$  matrix of explanatory variables that comprise the systematic, spatially-varying mean of the geostatistical process,  $\boldsymbol{\epsilon}$  is an  $(n \times 1)$  vector of error terms ( $\boldsymbol{\epsilon} \sim N(0, \tau^2)$ ) and  $p$  is the number of parameters that comprise the mean.

At any single location  $x$ :

$$Y(x) = \mu + \epsilon(x) \quad (3.11)$$

where  $Y$  is the value of the process at location  $x$ ,  $\mu$  is the mean and  $\epsilon(x) \sim N(0, \tau^2)$ .

Diggle and Ribeiro (2007) define a Gaussian spatial process  $T(x)$  to have mean  $\mu$ , variance  $\sigma^2$  and correlation function  $\rho(u) = \text{Corr}(T(x), T(y))$  where  $u = \|y - x\|$  for

any two locations  $x$  and  $y$  and where  $||.||$  denotes Euclidean distance.

Two important assumptions often made when fitting models to spatial data are that the underlying stochastic process is both stationary and isotropic. First-order, strong or strict stationarity is a largely theoretical concept and difficult to prove in practice. Sherman (2011) defines first-order stationarity as:

$$P(Z(x_1) \leq a_1, \dots, Z(x_k) \leq a_k) = P(Z(x_1 + h) \leq a_1, \dots, Z(x_k + h) \leq a_k) \quad (3.12)$$

for all shift vectors  $h$ . A more practical version of this concept is second-order or weak stationarity, which Sherman (2011) describes as having the following two properties:

- $E(Z(x)) = \mu$  for all locations
- $\text{Cov}(Z(x), Z(y)) = \text{Cov}(Z(x + h), Z(y + h))$  for any two locations  $x$  and  $y$  and all spatial shifts  $h$

The first property constrains the mean to be constant across space. The second property constrains the covariance function to depend only upon the spatial lag between any two locations  $x$  and  $y$  (i.e.  $u = y - x$ ) and not on the locations themselves (Sherman, 2011).

A concept related to second-order stationarity is that of intrinsic stationarity, which has similar properties:

- $E(Z(x) - \mu) = 0$
- $\text{Var}(Z(x + h) - Z(x)) = 2\gamma(h)$  for all shifts  $h$

This formulation is identical to that for second-order stationarity (Sherman, 2011)

with  $2\gamma(h)$  being called the variogram.  $\gamma(h)$  is known as the semi-variogram but is often simply referred to as the variogram also.

A stationary process is also isotropic if it is directionally-invariant i.e.  $\gamma(y - x) = \gamma(\|y - x\|)$  where  $\|\cdot\|$  denotes Euclidean distance (Diggle and Ribeiro, 2007). This means that the covariance between any two locations  $x$  and  $y$  depend only upon the distance between them and not on direction.

A geostatistical model assuming stationarity and isotropy can be used to make inference about a spatial process and such inference can be used to predict at unobserved locations, as aforementioned. Sherman (2011) states that the ‘best’ prediction of an unobserved location  $Z(x_0)$ , based on the observations  $Z(x_1), \dots, Z(x_n)$ , should typically have the lowest mean squared error i.e. the minimisation of:

$$\text{MSE}(\hat{Z}(s_0)) = E((\hat{Z}(s_0) - Z(x_0))^2) \quad (3.13)$$

## Variogram

The variogram of a geostatistical process is a widely-used alternative to covariance functions when assessing the presence of spatial correlation within residuals after a model has been fit.

Diggle and Ribeiro (2007) define the semi-variogram between two spatial locations  $x$  and  $y$  as follows:

$$\gamma(x, y) = \frac{1}{2} \text{Var}(T(x) - T(y)) \quad (3.14)$$

where the variogram,  $2\gamma(x, y)$ , would be scaled appropriately and where  $T(x)$  and  $T(y)$  define the spatial processes at locations  $x$  and  $y$ .

Diggle and Ribeiro (2007) also define the expanded version of Equation 3.14 in terms of variances and covariances:

$$\gamma(x, y) = \frac{1}{2} \text{Var}(T(x)) + \text{Var}(T(y)) - 2\text{Cov}(T(x), T(y)) \quad (3.15)$$

Figure 3.1 shows the theoretical empirical semi-variogram.

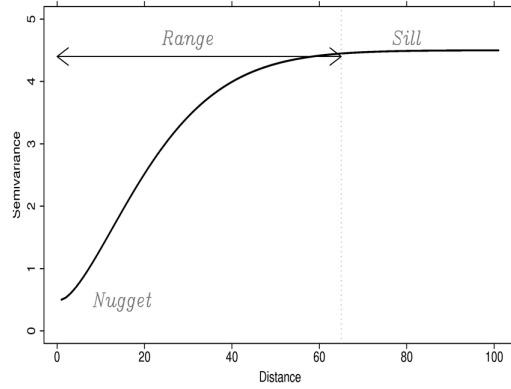


Figure 3.1: Theoretical semi-variogram (University of Edinburgh, 2015)

Diggle and Ribeiro (2007) state that most variograms are increasing, monotonic functions with four components: the range, the nugget, the sill and the partial sill (not shown in Figure 3.1).

The range is defined as the distance,  $t$ , where the semivariance crosses the sill, meaning that all subsequent points are effectively spatially-uncorrelated. The nugget is a “discontinuity at the origin of the variogram” (Diggle and Ribeiro, 2007) and the value of the semivariance as the distance between any two points  $x$  and  $y$  approaches zero (i.e.  $t = |y - x| \rightarrow 0$ ). It generally is seen as quantifying measurement error. The sill is an asymptote and represents the overall variance of the process. The partial sill is the difference between the sill and the nugget.

Determining whether spatial correlation exists or not is far from an exact science but

one can be assisted by the use of Monte Carlo envelopes. Analogous to confidence bands or intervals, Monte Carlo simulation is used to predict the behaviour of the variogram if the errors were truly independent. If the points in the variogram (especially the first few) trespass one or both of these envelopes, then it can be inferred that spatial correlation is most likely present. If the variogram appears to show evidence of spatial correlation in the model residuals then this has to be accounted for by fitting a variogram model. This involves modelling the correlation structure in the residuals as best as can be achieved.

A very general class of correlation models that are widely used in spatial analyses are the Matérn correlation functions (Matérn, 1960). Diggle and Ribeiro (2007 – derived from Matérn (1960)) define the general Matérn correlation function as follows:

$$\rho(u) = \left\{ 2^{\kappa-1} \Gamma(\kappa)^{-1} \left( \frac{u}{\phi} \right)^{\kappa} K_{\kappa} \left( \frac{u}{\phi} \right) \right\} \quad (3.16)$$

where  $u$  is the distance between two points,  $\Gamma$  is the gamma function,  $K_{\kappa}$  is a modified Bessel function of order  $\kappa$ ,  $\phi > 0$  is a scale parameter with the dimensions of distance and  $\kappa > 0$  is a shape parameter (Diggle and Ribeiro, 2007).

When  $\kappa = 0.5$ , the Matérn correlation function reduces to an exponential covariance structure, one of the most widely-used functions to model spatial correlation i.e.

$$\rho(u) \rightarrow \exp \left( -\frac{u}{\phi} \right) \quad (3.17)$$

### 3.1.3 Model Comparison

Several automatic techniques exist in order to allow competing models to be compared. One of the most common techniques is Akaike's Information Criterion or AIC (Akaike, 1974), which is defined by Fahrmeir et al. (2013) as:

$$\text{AIC} = -2l(\hat{\theta}) + 2p \quad (3.18)$$

where  $l$  is the log-likelihood of the  $p$ -dimensional parameter vector  $\hat{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T$ . The second term ( $2p$ ) penalises complex models. Models with the smallest AIC are favoured (Fahrmeir et al., 2013).

AICc (second-order or ‘corrected’ Akaike’s Information Criterion) is a variant of AIC and is less likely to result in overfitting due to extra constraints on parameters. It was originally proposed by Sugiura (1978) for linear models only but was extended to other types of statistical models by Hurvich and Tsai (1989). Burnham and Anderson (2002) strongly recommend the use of AICc, as opposed to AIC, when the sample size is particularly small or if the number of parameters is large. Nevertheless, AICc will converge to AIC when the sample size is suitably large.

### 3.1.4 Random Forests

The statistical methods discussed previously are typically only suitable for continuous response variables and applying them to non-continuous data would be difficult. In cases of binary or ordinal data, such as the macroinvertebrate data introduced in Chapter 2, alternative approaches should be considered.

Random forests (Breiman, 2001) are an extension of classification and regression trees. A classification tree gives a predicted classification for a set of categorical predictors and a random forest of  $k$  trees selects the classification which obtains the most votes over all the trees in the forest. Hence, it will choose the classification with the most votes in its favour at each prediction location. Like single classification trees, random forests do not make any distributional assumptions and do not overfit due to the Law of Large Numbers (Breiman, 2001).

A methodological pre-requisite is bootstrap aggregation or ‘bagging’ (Breiman, 1996), a key component of the random forest algorithm. It is described by Cutler (2010) as follows:

- A bootstrap sample is drawn at random with replacement from the data set
- Some observations will be sampled multiple times, whilst others will not be included (‘out-of-bag’)
- The process should reduce variance but has limited effect on bias, therefore it is most appropriate for classification trees where variance is typically high but bias is low

The random forest algorithm is detailed as follows (Breiman, 2001; Cutler, 2010):

- $k$  classification trees are constructed (large enough  $k$  will lead to convergence)
- Each tree is constructed from an independent bootstrap sample using the ‘bagging’ method
- $m$  variables are chosen at random from the  $M$  possible covariates in the model independently for each node in the tree
- The best split on the  $m$  variables is selected
- Classification trees are allowed to grow to maximum depth and are not pruned
- The  $k$  trees are averaged to obtain predictions for new data

Cases in the training data will be excluded from the bootstrap sample for about one third of the trees, termed ‘out-of-bag’ cases (Breiman, 2001). The out-of-bag error rate quantifies the effectiveness of the random forest model at predicting new data and is usually accompanied in statistical analyses by a classification or ‘confusion’ matrix made from the fitted random forest on the out-of-bag data (Breiman, 2001).

Prediction errors between classes will be highly unbalanced if the number of observations in each category is unbalanced (Breiman, 2001). More frequently-observed



classes will typically have lower prediction error whilst less frequently-observed classes will have higher prediction error. This is a result of the random forest attempting to minimise the overall error rate (Breiman, 2001).

## 3.2 Modelling Continuous Environmental Determinands

### 3.2.1 Soil Quality Models

The modelling process for all five metals was very similar. Hence, only one (lead) is given in depth here with corresponding results for the other four metals being displayed in Appendix B. A linear model consisting of three explanatory variables: the easting coordinate, the respective northing coordinate and an interaction term between these two main effects was first considered. The lead concentration was also log-transformed. The model was formulated as follows:

$$\log(y_{ij}) = \mu + \alpha s_i + \beta s_j + (\alpha\beta)s_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad (3.19)$$

where  $y_{ij}$  refers to the lead concentration at easting  $i$  and northing  $j$ ,  $\mu$  is an intercept term,  $\alpha s_i$  refers to the easting coordinate,  $\beta s_j$  refers to the northing coordinate,  $(\alpha\beta)s_{ij}$  refers to the interaction between the two main effects and  $\epsilon_{ij}$  refers to the error associated with the lead concentration at location  $s_{ij}$ .

However, due to the curvature observed in Figure 2.19, the model was refitted using an isotropic smooth for easting and northing. The model was formulated as follows:

$$\log(y_{ij}) = \mu + s(\gamma_{ij}) + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad (3.20)$$

where  $y_{ij}$  refers to the lead concentration at easting  $i$  and northing  $j$ ,  $\mu$  is an intercept term,  $\gamma_{ij}$  refers to the interaction between the two main effects,  $s$  defines the

variable to be a smooth function and  $\epsilon_{ij}$  refers to the error associated with the lead concentration at location  $(i, j)$ . The results of the flexible regression model, fitted using REML, are displayed in Table 3.1.

Table 3.1: Results of flexible regression model for lead

<b>Parametric Coefficient</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>t-value</b>	<b>p-value</b>
Intercept	4.821	0.016	300	< 0.05
<b>Smooth Coefficient</b>	<b>Effective Degrees of Freedom</b>	<b>Reference Degrees of Freedom</b>	<b>F-value</b>	<b>p-value</b>
Easting : Northing	17.05	22.03	17.65	< 0.05

The smooth spatial term was highly statistically-significant (p-value < 0.05). The linear and flexible regression models were compared using AICc, which confirmed that the latter was the preferred model. The smoothing parameter (i.e. the basis or number of knots), chosen using GCV, was chosen to be 110. This ensures that the  $k$ -index, which assesses residual variance, was at least 1. A  $k$ -index value lower than 1 means that a pattern in the residual variance may still remain, especially if the number of knots is close to the value for the effective degrees of freedom (Wood, 2006).

Figure 3.2(a) shows the assumption that the error term has constant variance may not be satisfied as the residuals appear to be projecting outwards as the fitted values increase. However, the assumption that the errors have mean zero does not appear to be an issue and neither does the assumption of normality, assessed in Figure 3.2(b). Figures 3.2(c) and 3.2(d) appear to show that the variability, if any, in the Easting and Northing variables has been accounted for by the model, hence, no evident pattern in either plot.

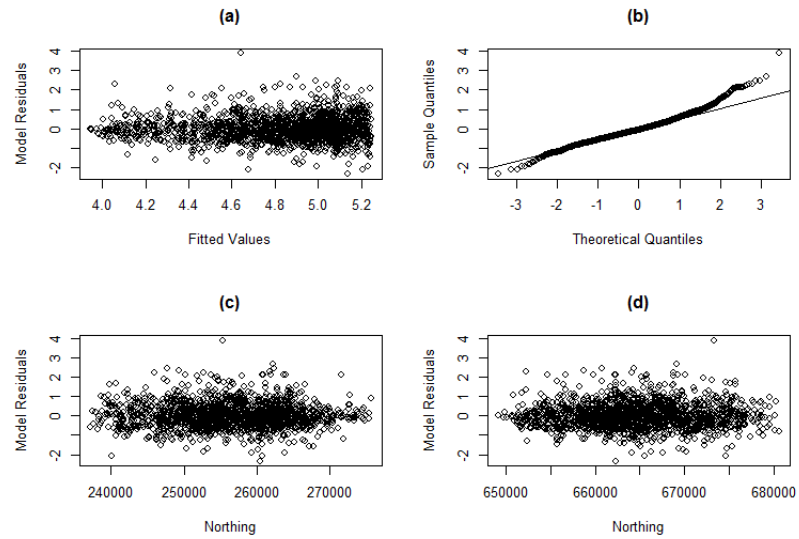


Figure 3.2: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

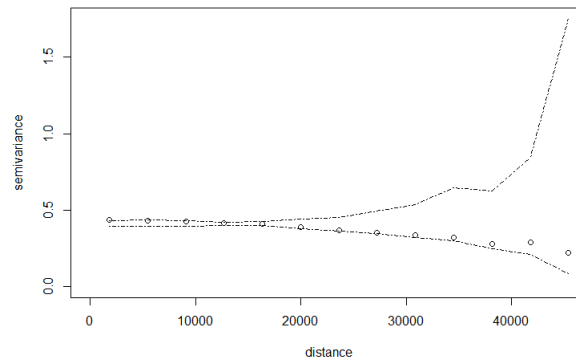


Figure 3.3: Variogram for lead flexible regression model residuals

The variogram in Figure 3.3 does not show evidence of significant spatial correlation in the model residuals as the variogram appears to be almost entirely within the bounds of the Monte Carlo envelopes, which test for such correlation. Therefore, it is not necessary to refit the model to account for any spatial correlation. Figure 3.4 shows how the lead concentration varies according to the fitted model. It appears

that the highest concentrations of lead appear to be in the highly-populated and urbanised areas of Glasgow City and that these concentrations gradually decrease as one moves out to more rural areas.

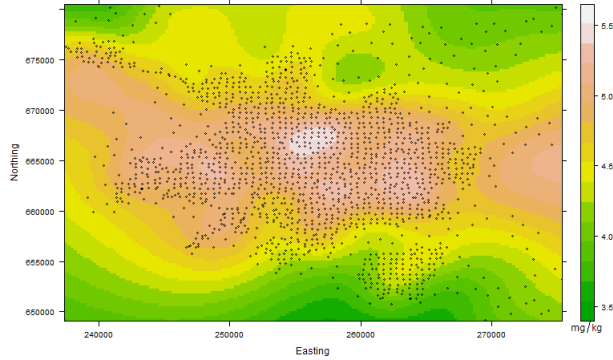


Figure 3.4: Prediction surface for flexible regression model for log-transformed lead concentration (mg/kg) with sampling locations

### 3.2.2 Air Quality Models

The geostatistical modelling procedure for the air pollutants was largely similar to that for the soil metals but model residuals here exhibited spatial correlation which had to be accounted for. The models described below were carried out on aggregated annual mean values for 2011 only. The modelling procedures and results for other years are similar and are not displayed here. Likewise, due to similarity, only the results for nitrogen dioxide are displayed here with results for other air pollutants being provided in Appendix B.

Firstly, a linear model consisting of an easting term, a northing term and an interaction term, identical to that defined mathematically in Equation 3.19 was fit. All three variables were highly-significant (all p-values  $< 0.05$ ). However, after the consideration of diagnostic plots (which showed that evident patterns remained in the model residuals) and the fact the curvature was present in the data (Figure 2.11), a flexible regression model was then fit in a manner identical to that in Equation 3.20 using REML.

Table 3.2: Results of flexible regression model for  $NO_2$ 

<b>Parametric Coefficient</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>t-value</b>	<b>p-value</b>
Intercept	1.936	0.0047	416	< 0.05
<b>Smooth Coefficient</b>	<b>Effective Degrees of Freedom</b>	<b>Reference Degrees of Freedom</b>	<b>F-value</b>	<b>p-value</b>
Easting : Northing	28.67	28.99	620.5	< 0.05

The number of knots was chosen to be 650, the first multiple of 25 which resulted in a  $k$ -index value which exceeded the value 1, meaning that there was unlikely to be any missed pattern in the residual variance. The number of knots was required to be much larger than that used for the soil quality models due to the larger amount of data. Whereas the soil quality models were fitted from approximately 1,600 data points, the air quality models were typically fitted from approximately 4,000 data points. A large number of knots was required to satisfy the criterion that the  $k$ -index should be at least 1. A test to compare the fits of the linear and flexible regression models was conducted using AICc, which again confirmed that the latter model was a better fit. The model was then further assessed using diagnostic plots.

The residual plots displayed in Figure 3.5 were much more difficult to make inference from than those for the soil metals. Figure 3.5(a) shows that the assumption that the errors had mean zero appeared reasonable but that such errors also having constant variance was difficult to confirm. A very heavy upper tail was present in the normal Q-Q plot in Figure 3.5(b), likely a result of the failure of the log-transformation to provide a normally-distributed response variable (see Figure 2.2). Figures 3.5(c) and 3.5(d) do not show much evidence of obvious patterns remaining the respective Easting and Northing variables but, again, are more difficult to make inference from than previous residual plots.

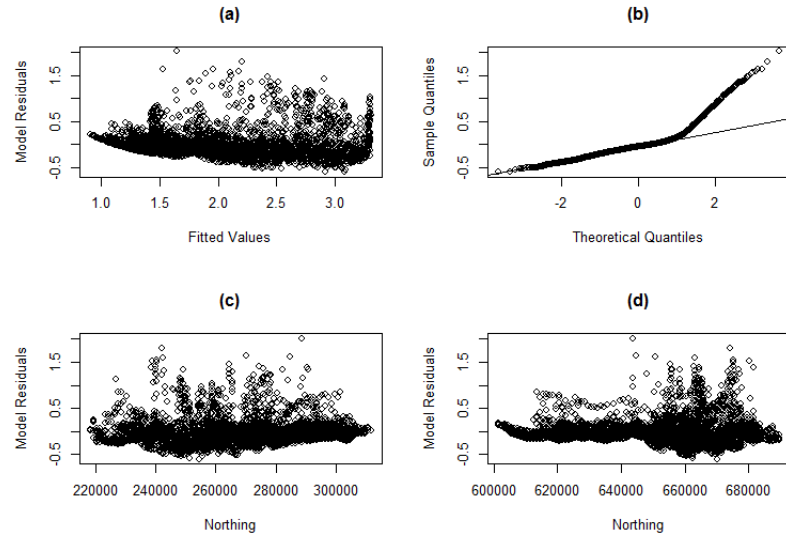


Figure 3.5: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

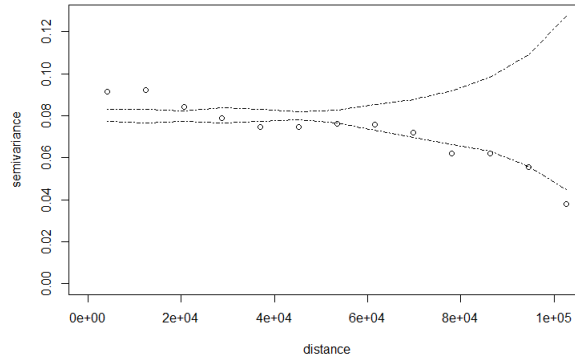


Figure 3.6: Variogram for  $NO_2$  additive model residuals

Figure 3.6 shows clear evidence of spatial autocorrelation in the residuals, as the points are not contained within the Monte Carlo envelopes. Hence, the model was refitted to include an exponential covariance structure to account for this. Finally, the model fit was visualised in terms of plotting the model surface against Easting and Northing, shown in Figure 3.7. Figure 3.7 focusses on what will be the region

of interest for constructing the composite indices in later chapters (i.e. the area shown in Figure 1.2). The results show largely what one would expect from the distribution of nitrogen dioxide, namely that very high concentrations would be observed in highly-populated areas and much lower values in less-populated areas.

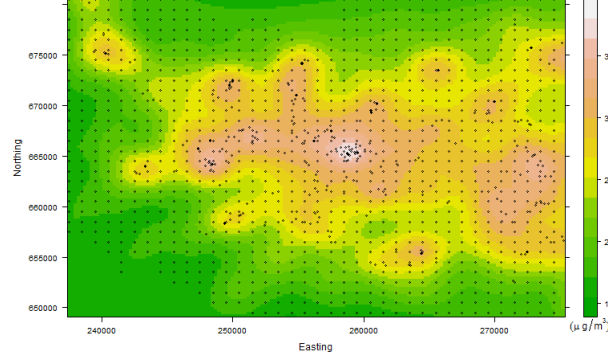


Figure 3.7: Prediction surface for flexible regression model for log-transformed  $NO_2$  concentration ( $\mu\text{g}/\text{m}^3$ ) with sampling locations

### 3.2.3 Water Quality Models for Chemical Determinands

For the five water chemistry determinands (ammonium, copper, dissolved oxygen, pH and soluble reactive phosphorus), the modelling process followed that for the air and soil quality models. However, a three-level categorical factor was also included to allow the use of river catchment information. Although the isotropic smooth remains, the addition of another covariate extends this model to an additive model. Each data point is located within either the River Clyde, River Kelvin or River Cart catchments. The modelling process will only be detailed for soluble reactive phosphorus here with similar results for the other four determinands being given in Appendix B. The model was formulated as:

$$\log(y_{ij}) = \mu + s(\gamma_{ij}) + \delta_k + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad (3.21)$$

where  $y_{ij}$  refers to the phosphorus concentration at easting  $i$  and northing  $j$ ,  $\mu$  is an

intercept term,  $\gamma_{ij}$  refers to the interaction between the two main effects,  $s$  defines the variable to be smooth function,  $\delta_k$  is a factor where  $k = 1$  (Clyde), 2 (Kelvin), 3 (Cart) and  $\epsilon_{ij}$  refers to the error associated with the phosphorus concentration at location  $(i, j)$ . The results of the flexible regression model, again fitted using REML, are displayed in Table 3.3.

Table 3.3: Results of flexible regression model for phosphorus

<b>Parametric Coefficient</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>t-value</b>	<b>p-value</b>
Intercept	-2.89	0.11	-26.23	< 0.05
Catchment Factor – River Kelvin	-1.55	0.33	-4.75	< 0.05
Catchment Factor – River Cart	-0.37	0.37	-0.99	0.33
<b>Smooth Coefficient</b>	<b>Effective Degrees of Freedom</b>	<b>Reference Degrees of Freedom</b>	<b>F-value</b>	<b>p-value</b>
Easting : Northing	4.03	5.38	12.75	< 0.05

Table 3.3 shows that the isotropic smooth is statistically-significant (p-value < 0.05) and there is a significant difference between phosphorus concentrations measured in the River Clyde and the River Kelvin. However, there is no evidence of a difference between measurements in the River Clyde and the River Cart (p-value > 0.05). The number of knots was chosen to be 15 for all five water chemistry models, which was much lower than the number of knots required for the air and soil quality models. However, this was due to the water data being relatively sparse in a spatial sense when compared to the air and soil data sets.

Table 3.4 displays the results of an analysis of variance to determine if the river catchment factor is statistically-significant. Table 3.4 shows that the catchment variable is highly-significant and should be kept in the model. Model diagnostics were then considered to assess its suitability.



Table 3.4: Results of analysis of variance for flexible regression model

Parametric Term	Degrees of Freedom	F-value	p-value
Catchment Factor	2	11.51	< 0.05

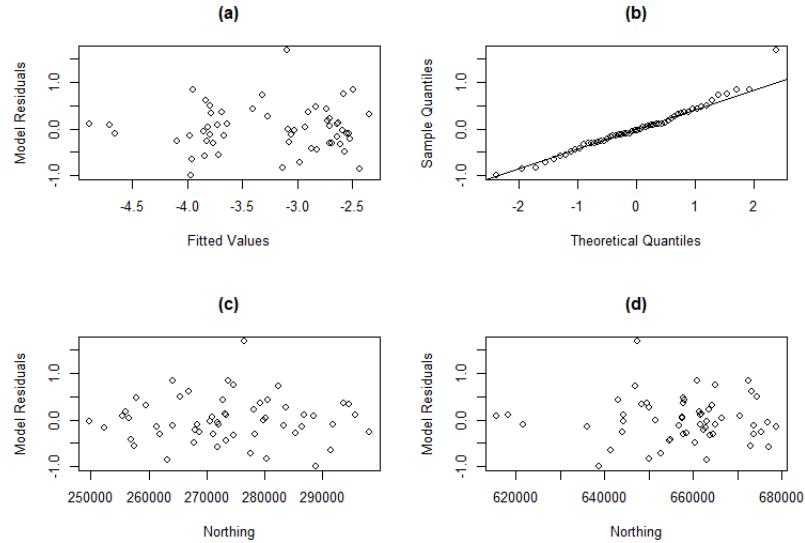


Figure 3.8: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

Figure 3.8(a) shows the model fitted values versus the residuals. It is difficult to make inference here, since there is significantly fewer data points than for the air and soil models. However, an obvious pattern does not appear to be present and the values are clustered around mean zero, albeit with one outlier. Figure 3.8(b) suggests that the assumption of normality is satisfied, again with one outlier. No patterns are visible in Figures 3.8(c) and 3.8(d), suggesting that the variability in the spatial coordinates has been accounted for by the model.

Figure 3.9 does not show evidence of any spatial correlation in the residuals.

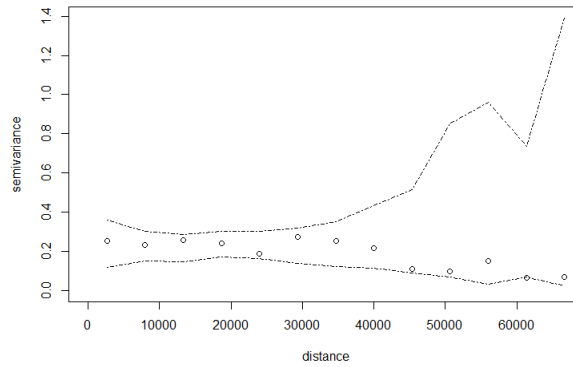


Figure 3.9: Prediction surface for flexible regression model for phosphorus concentration with sampling locations

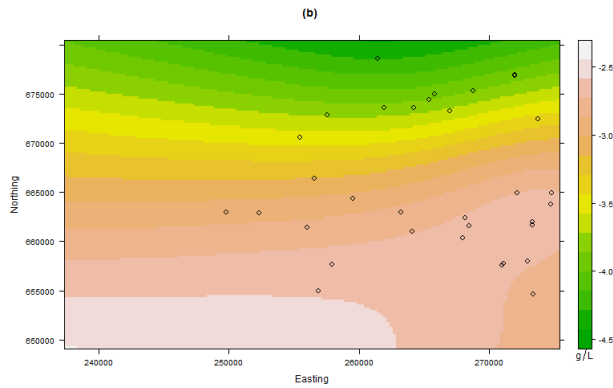


Figure 3.10: Prediction surface for flexible regression model for log-transformed phosphorus concentration (g/L) with sampling locations

Figure 3.10 shows the smooth surface fitted by the model. It appears that higher concentrations of soluble reactive phosphorus are found in the urban areas in the north-west of the range of the data (i.e. in and around Glasgow City). Lower concentrations are observed in rural areas.

### 3.3 Macroinvertebrate Analysis Using Random Forests

To ease the inclusion of macroinvertebrates into the composite index, the five-category ordinal scale was reduced to a two-category binary scale. The ‘bad’, ‘poor’ and ‘moderate’ categories were merged to create the ‘bad’ category and the ‘good’ and ‘high’ categories were merged to create the ‘good’ category. The data were modelled using a random forest classifier (Breiman, 2001) with three covariates (easting, northing and the river catchment factor, also used in the chemical determinand models). The forest consisted of 1,000 trees and one variable was automatically selected at each split in each tree. The resulting classification or ‘confusion’ matrix is displayed in Table 3.5.

Table 3.5: Classification matrix for random forest model

	<b>Bad</b>	<b>Good</b>	<b>Class Error</b>
<b>Bad</b>	7	9	0.56
<b>Good</b>	6	27	0.18

The overall prediction or ‘out-of-bag’ error rate was 30.61%, meaning that the random forest model is quite successful. Table 3.6 shows which variables were most important in the classification procedure, reported using a variable importance criterion based on Gini impurity (Breiman, 2001; Breiman and Cutler, no date). Higher Gini importance indicates a more important variable, as determined by the random forest classification process.

Table 3.6: Variable importance for random forest model

<b>Variable</b>	<b>Gini Importance</b>
Northing	7.48
Easting	5.76
Catchment	2.14

From Table 3.6, it can be inferred that the spatial coordinates are the most important in classifying the macroinvertebrate data but the river catchment information is also important. The random forest was then used to predict macroinvertebrate classifications at each prediction location. The indicator for the macroinvertebrate data will be expressed as the probability of a data zone being in the ‘good’ category.

### 3.4 Prediction and Aggregation to Data Zone Level

The purpose of modelling the environmental quality data described in Chapter 2 is to predict at unobserved locations which can then be used to create environmental quality indices, an approach used by Lee et al. (2011), amongst others. This is required as the indices will be reported at data zone level and at least one observation for each indicator is needed in each data zone for this to be possible.

A uniform grid was created using the statistical software package R (R Core Team, 2014), which consisted of approximately 17,000 data points across the 1,748 data zones in the study region (see Figure 1.2). Each point is a prediction location. This was to ensure that there would be at least one prediction location within the boundaries of every data zone that comprised the study region.

However, some data zones (especially within Glasgow City) were so geographically small that there were no prediction grid locations that fell within their boundaries. For each of these data zones, a prediction location was chosen manually and was selected to be as close to the geographic centroid of the data zone as possible. Approximately sixty data zones out of 1,748 required this manual process.

After each model had been fitted, the prediction grid was used to determine the spatial process across the study region, with at least one predicted observation in each of the 1,748 data zones. Since the data were modelled on the logarithmic scale, the predicted data were subsequently transformed back by applying an exponential

transformation before they were aggregated to data zone level.

A simple exponential back-transformation may not necessarily be the most appropriate back-transformation for data that have been modelled on the logarithmic scale, although this is a common practice (Newman, 1993). Newman (1993) suggests simple exponential back-transformations may introduce some bias into predictions but this was not investigated here.

Afterwards, a predicted value for each data zone was computed by calculating the arithmetic mean of all predicted observations in each data zone (Bruno and Cocchi, 2002; Lee et al., 2011). As described above, some geographically-small data zones had only one predicted observation. However, the vast majority of data zones had multiple, if not dozens, of prediction locations within their boundaries. This process resulted in one predicted value for each variable in each data zone in the study region, which were then used to create the composite indices in Chapter 4 of this thesis.

It should be noted that the water data was much more spatially-sparse than the air and soil data and meant that predictions may suffer from a greater amount of statistical variability. Also, the data were only available along the watercourses of the three rivers that comprised the data set. However, it was necessary to predict using the water quality models across all of the data zones in order for the composite index to be constructed in a consistent manner. Nevertheless, it was deemed a reasonable assumption that waterbodies (whether they be rivers, streams, canals etc.) would be present in almost every data zone and that predicting across the study region would be satisfactory for the purposes of this thesis. However, access to more data would undoubtedly be beneficial in a future study.

## 3.5 Chapter Summary

This chapter has shown the models applied to the data from Chapter 2 with a view to using model predictions to create a composite environmental quality index for Greater Glasgow. This has been achieved in all but one case through a combination of flexible regression models consisting of isotropic smooths of the spatial coordinates alongside a geostatistical analysis involving the consideration of spatial correlation and variograms.

Appropriate smoothing parameters for the four air quality pollutants were chosen by finding the first multiple of 25 which would result in a  $k$ -index of at least 1, thereafter residual variation should be largely random. Smoothing parameters for the five water chemical determinands were chosen to be 20. Smoothing parameters for the soil determinands varied but were still chosen such that the  $k$ -index was at least 1.

The fits of the models were also assessed by using residual plots, including checking whether the residuals were normally-distributed, as well as verifying if the isotropic smooth performed better than a corresponding normal linear model using AICc.

The ordinal macroinvertebrate data were modelled in a completely different manner. To ease future work, the five-category ordinal scale was condensed into a two-category binary scale and the data were modelled by using a random forest. The probability of a ‘good’ result at each prediction location was then calculated. The random forest model was assessed by the consideration of its misclassification rate.

Finally, these models were used to predict at a large number of locations in the study region and these predictions were aggregated using the arithmetic mean. This resulted in a value for each indicator in all 1,748 data zones. These values will then be used to construct composite indices at data zone resolution in Chapter 4.

# Chapter 4

## Development of Composite Indices

This chapter details the construction of composite environmental quality indices for Greater Glasgow for the year 2011. ‘Pseudo-data’ predicted from the models shown in Chapter 3 were used to create these indices, since the original data from Chapter 2 were not available at a spatial resolution high enough for the aims of this project.

### 4.1 Methods

#### 4.1.1 Index Construction Steps

For the reasons outlined in Chapter 1, it is vitally important that a transparent and reasoned approach is taken when constructing a composite index. This thesis will reflect on the appropriate steps recommended by the OECD (2008 – see Table 1.2). Step 1 (outlining a theoretical framework) is described in Section 1.4. Steps 2 and 3 (data selection and management) are accounted for in Chapter 2. This chapter will focus on steps 4, 5 and 6, namely multivariate analysis and the normalisation, weighting and aggregation of indicators.

Figure 4.1 is a flow chart detailing the general steps required to create the final environmental quality index.

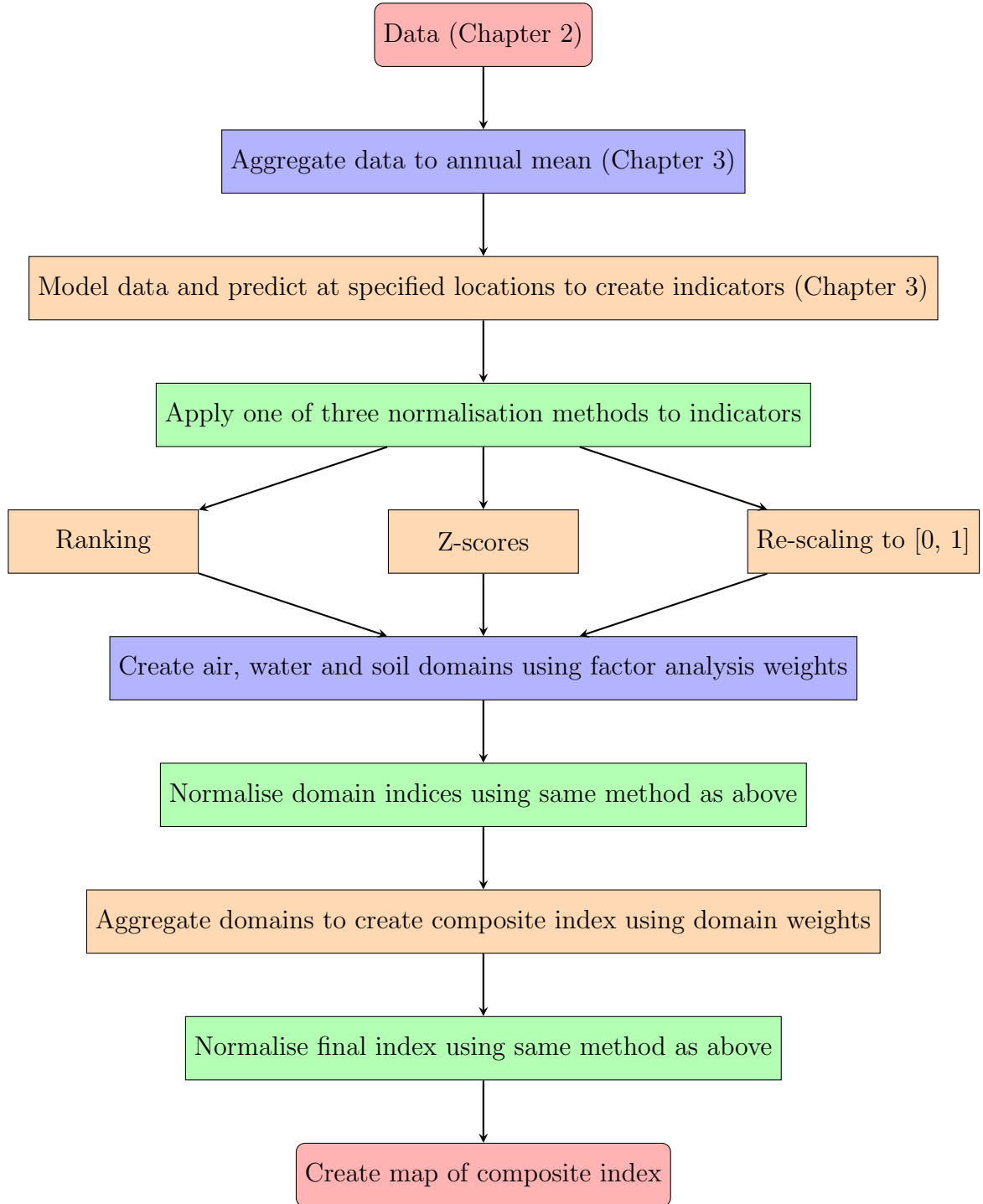


Figure 4.1: Flow chart detailing general steps in constructing composite indices



The three normalisation methods used in this thesis are described below and will be used to create three different composite indices, all measuring environmental quality and consisting of the same domains and indicators.

## **Ranking**

The ‘ranking’ method that will be used to normalise indicators will follow the methodology of the Scottish Index of Multiple Deprivation (Scottish Government, 2012 – see Figure 1.1). The steps that will be taken are outlined below:

1. Each indicator is ranked from 1 (best environmental quality) to 1,748 (worst environmental quality)
2. The indicators are weighted according to factor analysis and aggregated to create a domain score (i.e. an air domain, a water domain and a soil domain)
3. Each domain is normalised again using the same method as Step 1, exponentially-transformed (according to Equation 4.1) and aggregated together to create an environmental quality index (EQI) using domain weights from factor analysis
4. The domain scores are ranked again from 1 to 1,748

Step 1 normalises the indicator scores and ensures that all indicators are transformed onto the same scale prior to aggregation. Step 2 weights each indicator (within its respective domain) to avoid issues of double-counting and multicollinearity. It is not appropriate to aggregate indicators that are very highly correlated with each other without applying a weighting scheme (either arbitrary, through consultation with experts or through statistical methods such as factor analysis or principal components analysis). The indicator scores are summed together using the appropriate weighting. This creates an air quality index (AQI), a water quality index (WQI) and a soil quality index (SQI). These domains are again normalised to achieve the desired index scale.

SIMD (Scottish Government, 2012) recommends that the domain scores are sub-

jected to an exponential transformation to avoid high values in one domain being cancelled out by low values in another. This will occur as the domain ranks will otherwise be uniformly-distributed. The transformation is shown in Equation 4.1 (Scottish Government, 2012).

$$-23 \times \log\{1 - R \times [1 - \exp(-100/23)]\} \quad (4.1)$$

where  $\log$  is the natural logarithm and  $R$  denotes a data zone's rank within each of the three domains, scaled to the range  $[0, 1]$  i.e. ( $R = 1/1748$  for the data zone with the best environmental quality in each data zone and  $R = 1748/1748$  for the data zone with the worst environmental quality in each data zone).

The three transformed domains are now aggregated using pre-determined weights (drawn from the loadings of the aforementioned factor analysis). Finally, as described in Step 4, the final index is ranked again from 1 (data zone with best environmental quality) to 1,748 (data zone with worst environmental quality). This results in a ranked composite environmental quality index (EQI) consisting of an air, soil and water domain for the study region shown in Figure 1.2.

This method is advantageous because it is unaffected by outliers (OECD, 2008). However, it is important to recognise that, like SIMD (Scottish Government, 2012), data zones can only be compared relative to each other. Hence, it is not appropriate to conclude that the environmental quality of data zone which is ranked 500<sup>th</sup> is twice as good as that of the data zone which is ranked 1000<sup>th</sup>. The only conclusion that can be drawn is that data zone 500 has better environmental quality than data zone 1000.

### **Z-scores**

This second method uses z-scores (often called standardisation) in place of ranks to create the same composite environmental quality index. The steps taken are shown

below:

1. The z-scores of each air, soil and water determinand are taken
2. The normalised scores are weighted according to factor analysis and aggregated to create a domain score
3. The resulting domain scores are normalised using z-scores, weighted according to the results of the factor analysis and aggregated to create the EQI
4. The EQI is subject to a z-score transformation

A z-score for each data zone is calculated by subtracting the mean data zone concentration for the determinand in question and dividing by the standard deviation of all concentrations i.e.

$$z = \frac{x - \mu}{\sigma} \quad (4.2)$$

where  $z$  is the z-score for each indicator,  $x$ , and  $\mu$  and  $\sigma$  are the mean and standard deviation, respectively, for the indicator. This results in a mean of zero and a standard deviation of one (Nardo et al., 2005). However, Nardo et al. (2005) also observe that indicators with extreme values will have a greater effect on the resulting composite index. This may be appropriate in some situations where extremes are of interest but perhaps not in other cases.

An attractive feature of this method in the context of this thesis is that it highlights the data zones which are of the greatest concern (i.e. z-score  $> 0$ ) with regards to air, soil and water quality as well as the general environmental quality shown by the final EQI.

### **Re-scaling**

This third and final method involves re-scaling all indicators to the interval  $[0, 1]$ . This is achieved by subtracting the mean value from each indicator value and then

dividing by the range (i.e. the minimum value subtracted from the maximum value).

Creating an index from such normalised indicators follows the exact same methodology as that for the z-score method described above. An advantage of re-scaling over standardisation is that the resulting scale of the index is more easily interpretable. The index will vary between 0 and 1, with data zones closer to 1 having increasingly poorer environmental quality.

#### **4.1.2 Factor Analysis**

As aforementioned, it is unwise to combine indicators which are highly correlated with each other without addressing the issue. Correlation between indicators leads to double-counting and multicollinearity problems, which can be solved in a variety of ways. Nardo et al. (2005) and OECD (2008) provide details on a variety of possible approaches.

Some developers of composite indices simply allow each indicator to be equally-weighted, which may be appropriate in some situations. Others apply an arbitrary weighting system, usually in consultation with experts in the field or in line with an established precedent. However, a statistical approach allows for easier justification and a better defence to criticism from statisticians, as described by Saisana et al. (2005). In this thesis, factor analysis will be applied to the fourteen continuous variables to find a suitable weighting system for each domain and for the final EQI.

Since the macroinvertebrate data are not continuous, it is not appropriate for them to be included in a standard factor analysis using Pearson's correlation. Polychoric correlation, or a related method, would have to be used in place of the commonly-used Pearson's correlation coefficient to account for the presence of both continuous and categorical variables. Hence, the macroinvertebrate indicator was allocated an arbitrary weighting in line with its importance as part of the water quality component of the index. The indicator was set to receive a weighting of 25% within the

water quality domain. The factor analysis weights for the remaining five chemical determinands in the water quality index were scaled down appropriately to account for this.

Factor analysis is closely related to principal components analysis or PCA (Pearson, 1901). The aim of PCA is to transform a set of correlated variables into a new set of uncorrelated variables using the covariance or correlation matrix (Nardo et al., 2005). The resulting variables are linear combinations of the original variables and typically the first few principal components should account for the vast majority of the variation in the data, if there is enough intra-variable correlation to begin with. PCA is commonly used for dimension reduction, which is not its purpose of this thesis. Factor analysis is based on a particular model (Spearman, 1904) and can be used to analyse the intra-indicator correlations and identify which indicators explain the most variability in their respective domain.

Weightings can be obtained from squared factor loadings. These are calculated from the matrix of factor loadings after rotation, “given that the square of factor loadings represents the proportion of the total unit variance of the indicator which is explained by the factor” (OECD, 2008).

Nardo et al. (2005) define the general form of a factor analysis model:

$$\begin{aligned}
 x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + e_1 \\
 x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + e_2 \\
 &\vdots \\
 x_Q &= a_{Q1}F_1 + a_{Q2}F_2 + \cdots + a_{Qm}F_m + e_Q
 \end{aligned} \tag{4.3}$$

where  $x_i$  is a variable with mean zero and unit variance;  $a_{i1}, a_{i2}, \dots, a_{im}$  are factor loadings corresponding to the variable  $X_i$ ;  $F_1, F_2, \dots, F_m$  are  $m$  uncorrelated factors

and  $e_i$  are the  $Q$  factors which are independently and identically distributed with mean zero (Nardo et al., 2005).

Nardo et al. (2005) observe that it is common to use PCA to extract the first  $m$  principal components, often with the assistance of a scree plot (Cattell, 1966), and then use that number of factors in a subsequent factor analysis. The number of factors that are extracted is often arbitrary and dependent upon the data set in question. It is common to use a ‘stopping rule’, such as the Kaiser criterion (Kaiser, 1960), in which all factors with eigenvalues less than 1 are dropped (Nardo et al., 2005).

## 4.2 Domain Indices

This section will detail the development of an index for each of the three domains (air, soil and water) prior to creating an environmental quality index involving all three domains.

### 4.2.1 Air Quality Index

An air quality index (AQI) for 2011 was created using the three normalisation methods. The composite index consisted of  $NO_2$ ,  $NO_X$ ,  $PM_{10}$  and  $PM_{2.5}$ . Firstly, factor analysis was used to determine the weights for the index, shown in Table 4.1.

Table 4.1: Indicator weightings for air quality index (AQI)

Indicator	Approximate Weight
$NO_2$	21%
$NO_X$	26%
$PM_{10}$	26%
$PM_{2.5}$	26%

The factor analysis has not strayed far from equal weighting with regards to the air

quality indicators although, as Table 4.1 shows,  $NO_2$  has been weighted less heavily than the other three determinands. These weights will be used for all three AQI construction methods.

## Ranking

The SIMD (Scottish Government, 2012) ranking method outlined in Section 4.1.1 was then followed to create the air quality index, shown in Figure 4.2.

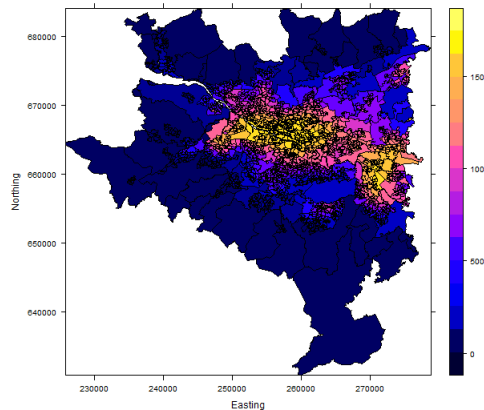


Figure 4.2: Air quality index (AQI) for 2011 using ranking method (1 = best air quality; 1748 = worst air quality) – contains Ordnance Survey data © Crown copyright and database right (2014)

Figure 4.2 confirms what was often seen in the exploratory analyses of the air quality determinands, namely that poorer air quality appears to be observed in highly-urbanised areas with better air quality being observed in rural areas with lower levels of population and traffic. In Figure 4.2, the areas of most concern appear to be Glasgow City and the highly-populated areas of North and South Lanarkshire to the south-east of the city.

## Z-scores

The z-score AQI shown in Figure 4.3 exhibits a similar spatial pattern to Figure 4.2 i.e. densely-populated data zones appear to suffer from poorer air quality with rural

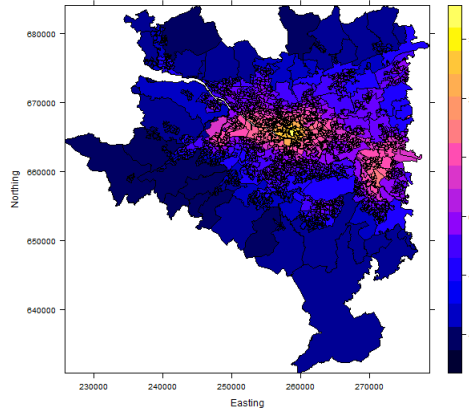


Figure 4.3: Air quality index (AQI) for 2011 using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014)

areas being of less concern. The spatial distributions of air quality are not identical to Figure 4.2 with the z-score method highlighting specifically Glasgow City centre as having the poorest air quality. The data zones covering the large towns of North and South Lanarkshire appear to exhibit better air quality according to this method.

### Re-scaling

Finally, the AQI constructed using the re-scaling normalisation approach is displayed in Figure 4.4

The results shown in Figure 4.4 appear very similar to those suggested in Figures 4.2 and 4.3, namely the worst-affected data zones tend to be in highly-urbanised areas.

### 4.2.2 Soil Quality Index

An almost identical approach was followed with regards to the soil quality index as that of the air quality index in Section 4.2.1. It should be reiterated, as mentioned in Chapter 2, that data were only available for one year (2001 – 2002) and, hence, all indices constructed in this chapter will assume that soil quality has remained largely constant since then. Firstly, the results of the domain factor analysis are shown in



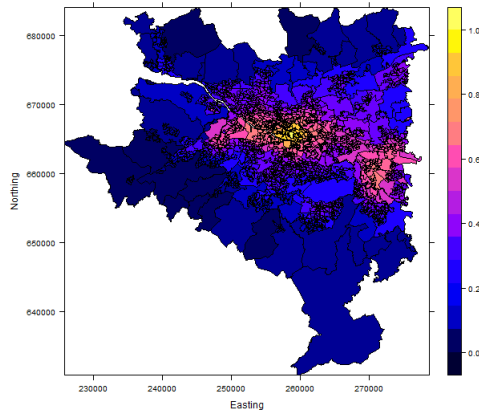


Figure 4.4: Air quality index (AQI) for 2011 using re-scaling method (0 = best air quality; 1 = worst air quality) – contains Ordnance Survey data © Crown copyright and database right (2014)

Table 4.2.

Table 4.2: Indicator weightings for soil quality index (SQI)

Indicator	Approximate Weight
Arsenic	29%
Chromium	24%
Lead	12%
Nickel	23%
Selenium	11%

Whilst factor analysis weightings for the air quality domain (Table 4.1) were broadly equal, there is an obvious departure from this for the soil determinand weightings. As Table 4.2 shows, arsenic, chromium and nickel receive a heavier weighting than lead or selenium. This is likely to produce a different SQI than one which more or less allowed equal weighting. The SQI was then constructed using the three normalisation methods.

## Ranking

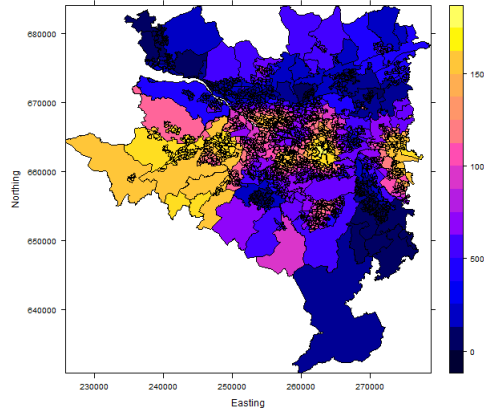


Figure 4.5: Soil quality index (SQI) using ranking method (1 = best soil quality; 1748 = worst soil quality) – contains Ordnance Survey data © Crown copyright and database right (2014)

Figure 4.5 shows the SQI following the SIMD (Scottish Government, 2012) method. Unlike the AQI, the spatial pattern of the index is not intuitive. As shown in Figures 4.2, 4.3 and 4.4, it is easy to imagine that poorer air quality would be present in highly-populated areas. However, this does not necessarily reflect the spatial distribution of soil quality. Figure 4.5 does highlight Glasgow City and other urbanised areas as generally having poorer soil quality than most rural areas but the local authority of Renfrewshire to the west of the study region is also highlighted as having particularly poor soil quality.

## Z-scores

Figure 4.6 exhibits a similar spatial pattern to Figure 4.5, meaning that the SQI is likely to be a good reflection of soil quality in the study region. Like the z-score AQI (Figure 4.3), Figure 4.6 again specifically highlights the areas with the poorest soil quality. Here, this is at the extreme west of the study region in the local authority of Renfrewshire.

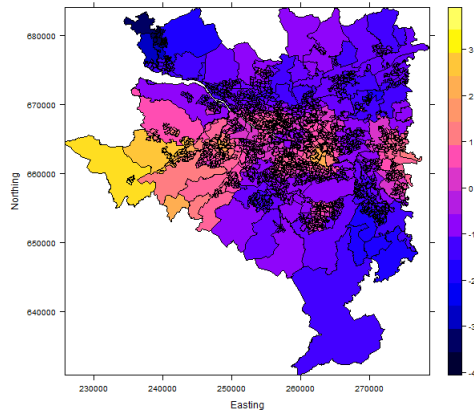


Figure 4.6: Soil quality index (SQI) using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014)

### Re-scaling

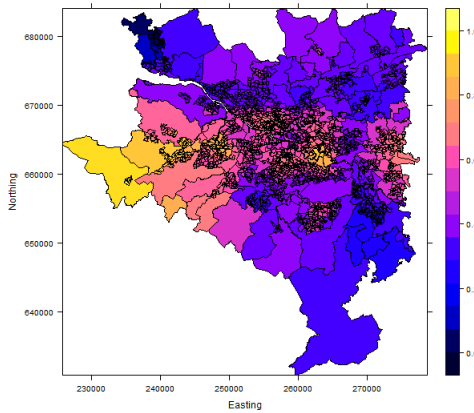


Figure 4.7: Soil quality index (SQI) using re-scaling method (0 = best soil quality; 1 = worst soil quality) – contains Ordnance Survey data © Crown copyright and database right (2014)

Figure 4.7 again shows similar results to Figures 4.5 and 4.6, implying that central Glasgow and the far-west of the study region suffer from the poorest soil quality.

### 4.2.3 Water Quality Index

Finally, the three normalisation methods were used to create a water quality index (WQI). Table 4.3 shows the weightings given to the five chemical determinands as generated from a factor analysis as well as their respective scaled weightings once the macroinvertebrate indicator had been taken into account, which was chosen to be 25%.

Table 4.3: Indicator weightings for water quality index (WQI)

Indicator	Approximate Weight (FA)	Scaled Weight
Ammonium	12%	9%
Copper	14%	11%
Dissolved Oxygen	7%	5%
pH	34%	26%
Soluble Reactive Phosphorus	32%	24%
Macroinvertebrates	–	25%

The factor analysis for the five chemical determinands gives particularly heavy weighting to pH and soluble reactive phosphorus in the WQI. The low weighting given to dissolved oxygen is particularly of note, perhaps due to lack of dependence on the spatial coordinates as shown in Table B10 in Appendix B. The low weighting is likely a reflection of this lack of variability in dissolved oxygen concentrations across the study region.

Two pre-construction alterations were considered for two of the chemical water quality indicators. For the arsenic, chromium, lead and soluble reactive phosphorus indicators, higher values means that environmental quality is deteriorating. However, since higher dissolved oxygen is generally seen as an indication of better water quality, the indicator was inverted. This meant that the direction of the dissolved oxygen indicator was consistent with that of the others in the water quality index.

Also, since the desired pH value of river water should be 7, meaning the water is neither acidic nor alkaline, it may be desirable to have the value zero be the baseline for the indicator. This can be achieved by subtracting 7 from all pH data zone values. However, since no data zone had a predicted pH value less than 7, this calibration is not necessarily required here.

## Ranking

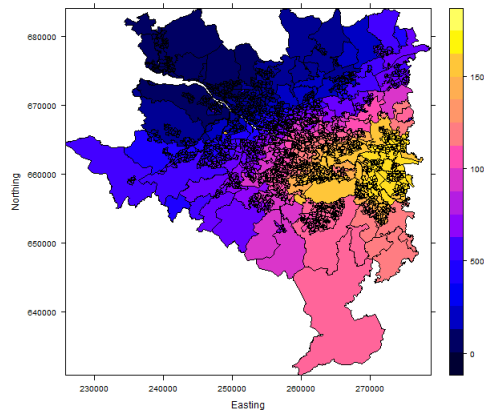


Figure 4.8: Water quality index (WQI) for 2011 using ranking method (1 = best water quality; 1748 = worst water quality) – contains Ordnance Survey data

© Crown copyright and database right (2014)

The WQI shown in Figure 4.8 appears to suggest that the poorest water quality is observed in the east of the study region in and around the major towns of Lanarkshire and the eastern borders of Glasgow.

## Z-scores

Figure 4.9 shows a very similar spatial pattern to Figure 4.8 in that the index suggests poorer water quality is observed in the urbanised east of the study region with improving water quality as one moves in a north-westerly direction.

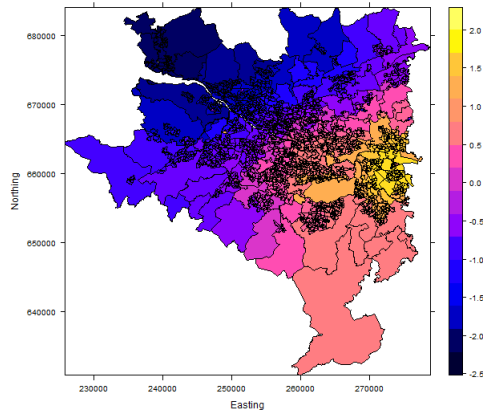


Figure 4.9: Water quality index (WQI) for 2011 using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014)

### Re-scaled

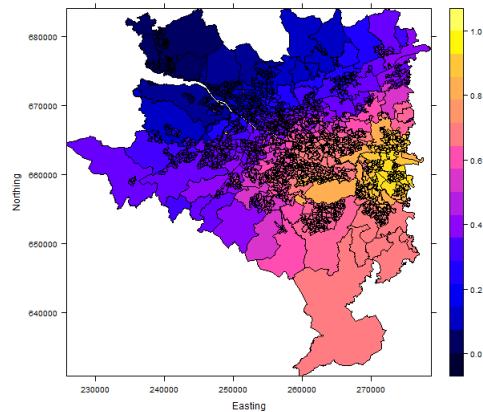


Figure 4.10: Water quality index (WQI) for 2011 using re-scaling method (0 = best water quality; 1 = poorest water quality) – contains Ordnance Survey data © Crown copyright and database right (2014)

Figure 4.10, the results of the WQI for the re-scaling normalisation method, suggests that the spatial pattern of water quality is very similar to Figures 4.8 and 4.9. Again, the data zones which suffer from the worst water quality appear to be in central Lanarkshire (the east of the study region).

## 4.3 Environmental Quality Index

This section outlines how the final environmental quality index (EQI) is created using the results of the air, soil and water domain indices in Section 4.2. The EQI will be developed using the three normalisation methods and will apply domain weightings from the same factor analysis used to extract the indicator weightings in Tables 4.1, 4.2 and 4.3. Table 4.4 shows the domain weightings for the EQI.

Table 4.4: Domain weightings for environmental quality index (EQI)

Domain	Approximate Weight
Air	38%
Soil	35%
Water	27%

### Ranking

Figure 4.11 shows the EQI created using the ranking method developed by SIMD (Scottish Government, 2012).

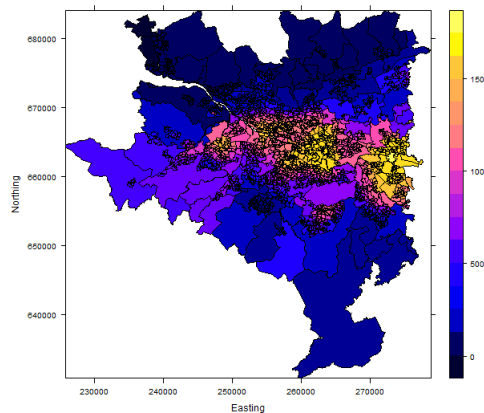


Figure 4.11: Environmental quality index (EQI) for 2011 using ranking method (1 = best environmental quality; 1748 = worst environmental quality) – contains Ordnance Survey data © Crown copyright and database right (2014)

Figure 4.11 suggests that the densely-populated Glasgow City and other highly-

urbanised areas, especially to the east of the study region, have the poorest environmental quality across the domains of air, soil and water. This is likely because of these areas' poor performance with regards to all three domains, especially air. It should also be noted that the rural north of the study region (largely comprising of the local authorities of East and West Dunbartonshire) appears to have much better environmental quality than the rural south. This is likely due to the fact that poorer soil quality was observed in the Renfrewshire area and poorer water quality in the south-east of the study region which has resulted in these areas having a lower overall score than more northern areas. The rural north generally performed well across all three domains, leading to the index suggesting that these data zones have the best environmental quality in Greater Glasgow.

### Z-scores

The EQI created using the z-scores method is shown in Figure 4.12.

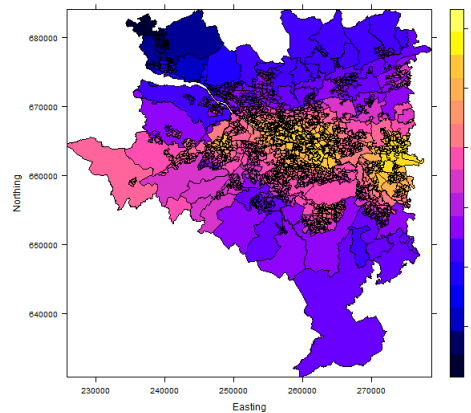


Figure 4.12: Environmental quality index (EQI) for 2011 using z-score method – contains Ordnance Survey data © Crown copyright and database right (2014)

A similar spatial pattern is observed in Figure 4.12 with regards to the corresponding EQI in Figure 4.11. Figure 4.12 again highlights highly-urbanised data zones as appearing to have the worst overall environmental quality. However, the z-score method has also shown striking differences in rural areas. The far-west of the study



region is coloured noticeably brighter when compared to other southern rural areas to the east, meaning that these data zones are performing worse with this method. This could be due to the z-score method again focussing on particular areas where environmental quality is particularly poor or that the soil domain is given greater weighting in the EQI than the WQI, which caused the south-east of the study region to perform worse under the ranking method. This disparity between the western and eastern sections of the rural south is evident when compared to Figure 4.11 where the southern data zones were generally coloured the same.

### Re-scaling

Finally, the EQI created using the re-scaling normalisation method is shown in Figure 4.13.

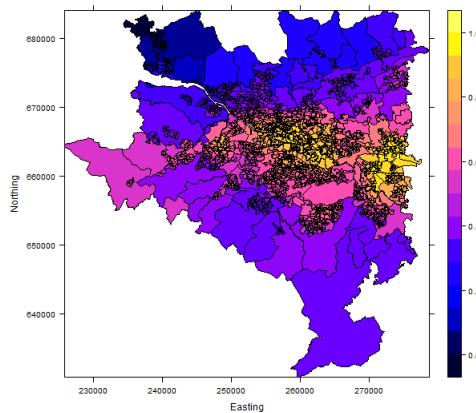


Figure 4.13: Environmental quality index (EQI) for 2011 using re-scaling method  
(0 = best environmental quality; 1 = worst environmental quality) – contains  
Ordnance Survey data © Crown copyright and database right (2014)

The data zone results displayed in Figure 4.13 largely exhibit the same spatial pattern to those in Figure 4.12. Central Glasgow and such major towns as Motherwell and Hamilton in central Lanarkshire typically exhibit the poorest overall environmental quality. The northern part of the study region, in particular the north-west, are suggested to have the best environmental quality.

## 4.4 Tabular Summary of Results

The results of the indices are displayed for eleven data zones in Tables 4.5, 4.6 and 4.7, for the ranking, z-score and re-scaling normalisation methods, respectively. The data zones were chosen according to the results of the index constructed using the ranking method and the factor analysis weights (Figure 4.11). The best and worst-performing data zones were chosen as well as those which corresponded to the 10<sup>th</sup>, 20<sup>th</sup>, . . . , 80<sup>th</sup> and 90<sup>th</sup> percentiles of the index. Figure 4.14 shows the spatial location of the selected data zones, which covers both urban and rural areas of the study region.

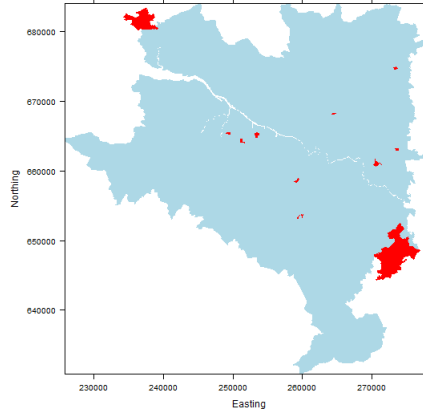


Figure 4.14: Location of eleven selected data zones in Greater Glasgow (those included are coloured red)

## 4.5 Chapter Summary

This chapter has used the ‘pseudo-data’, obtained from the models fitted in Chapter 3, to construct composite indices for the study region for the year 2011. It should again be noted that the soil data were only available for the year 2001 – 2002 and soil pollution levels were assumed to have remained constant since then, for the purposes of developing methods for this thesis. This was achieved through the procedure outlined in Figure 4.1, involving the normalisation of indicators and then

Table 4.5: Ranked index values for eleven selected data zones (1 = best environmental quality; 1748 = worst environmental quality)

<b>Data Zone</b>	<b>AQI</b>	<b>SQI</b>	<b>WQI</b>	<b>EQI</b>
S01003029	314	717	1459	700
S01003375	1559	1055	640	1224
S01003588	1176	918	914	1049
S01004680	1625	715	1694	1573
S01004716	1519	1592	1688	1748
S01004896	916	458	592	525
S01005258	856	1059	633	875
S01005288	1454	1612	390	1399
S01005741	63	54	1274	176
S01005835	241	481	1037	350
S01006286	77	6	24	1

Table 4.6: Z-score index values for eleven selected data zones

<b>Data Zone</b>	<b>AQI</b>	<b>SQI</b>	<b>WQI</b>	<b>EQI</b>
S01003029	-0.926	-0.307	1.068	-0.255
S01003375	1.386	0.266	-0.460	0.719
S01003588	0.396	0.071	0.041	0.269
S01004680	-0.309	-1.595	0.053	-0.965
S01004716	1.241	1.218	1.682	1.948
S01004896	-0.006	-0.091	-0.397	-0.616
S01005258	-0.110	0.274	-0.481	-0.105
S01005288	0.982	1.248	-0.808	0.867
S01005741	-1.689	-1.579	0.900	-1.389
S01005835	-1.113	-0.721	0.524	-0.778
S01006286	-1.587	-3.430	-2.214	-3.474

Table 4.7: Re-scaled index values for eleven selected data zones (0 = best environmental quality; 1 = worst environmental quality)

<b>Data Zone</b>	<b>AQI</b>	<b>SQI</b>	<b>WQI</b>	<b>EQI</b>
S01003029	0.253	0.475	0.790	0.604
S01003375	0.691	0.585	0.415	0.757
S01003588	0.499	0.554	0.519	0.676
S01004680	0.720	0.507	0.927	0.930
S01004716	0.651	0.726	0.916	1.000
S01004896	0.425	0.403	0.408	0.515
S01005258	0.406	0.582	0.414	0.598
S01005288	0.602	0.731	0.335	0.752
S01005741	0.113	0.281	0.726	0.403
S01005835	0.221	0.402	0.647	0.495
S01006286	0.127	0.021	0.004	0.000

aggregating them together using weights from factor analysis. Domain indices were first created for air quality, soil quality and water quality, individually before the process was repeated in order to aggregate the domains into a final environmental quality index. The composite index was created using three different types of normalisation methods: ranking, z-scores and re-scaling to  $[0, 1]$ .

In general, the three normalisation strategies resulted in a similar spatial process being suggested for air, soil, water and overall environmental quality. For the air domain, all three indices suggested that the worst air quality appeared to be observed in highly-urbanised areas, especially in the centre of the city of Glasgow. For the soil domain, the indices implied that poorer soil quality was typically observed in the far-west of the study region. The indices for the water domain suggested that poorer water quality was found in the urbanised eastern areas of the study region.

Finally, the overall environmental quality indices suggested that the poorest environmental quality was observed in urban areas (such as the city of Glasgow and the major towns of Lanarkshire to the immediate east of Glasgow) with better results being found in rural areas, particularly in the data zones to the far north.

The robustness of these composite indices will be assessed in Chapter 5.

# Chapter 5

## Assessment of Composite Indices

### 5.1 Introduction

This chapter will concentrate on assessing the composite environmental quality indices developed in Chapter 4 by investigating:

- How statistical uncertainty within the indices can be quantified?
- How the indices can be used to gain insight into how environmental quality changes over a period of five years?

Assessing the level of uncertainty with the indices is essential as throughout the development of a composite index, several subjective choices have to be made (OECD, 2008) such as:

- Choosing appropriate data to create the index itself
- The choice of normalisation method
- The selection of individual weights for each indicator
- The choice of aggregation model

amongst others. Due to the stochastic nature of environmental systems, a statement of confidence should be given for each response (Saltelli, 2000). In practical terms

with regards to this thesis, an interval of uncertainty should be calculated for each data zone index value. This uncertainty-based assessment of composite indices is still relatively uncommon in the literature, particularly those constructed on a ranked scale.

Several statistical techniques have been developed to assess the impact of these subjective judgements and are often grouped into ‘uncertainty and sensitivity analyses’. Uncertainty and sensitivity analyses are two distinct but very closely related processes. Saisana et al. (2005) state that uncertainty analysis typically focusses on “how uncertainty in the input factors propagates through the structure of the composite index and affects the values of the composite index”, whilst sensitivity analysis quantifies “how much each individual source of uncertainty contributes to the output variance”. In other words, uncertainty analysis attempts to estimate the level of uncertainty in inference whilst sensitivity analysis attempts to pinpoint which parts of the index are largely responsible for this uncertainty (Saltelli and Annoni, 2010).

Saltelli and Annoni (2010) state that uncertainty and sensitivity analyses are complimentary, not alternatives to one another. Saisana et al. (2005) also observe that examples of uncertainty analysis are more commonly found in the literature and uncertainty analysis will be the primary focus of this chapter.

The main advantage of undertaking an uncertainty analysis when constructing a composite index is that it may improve the index’s inferential value and lead to more defensible results (Dobbie and Clifford, 2015). This is important due to the scepticism that exists within the statistical community towards composite indices as a whole (Saisana et al., 2005).

Saisana et al. (2005) observe that propagating uncertainties leads to an index being a distribution of values, rather than a simple number. In many examples in the literature, this almost certainly leads to overlap between different partitions of the

study region, whether they be countries, administrative districts or data zones, as in this thesis. This could be seen by some as a potential downside to carrying out an uncertainty analysis on a composite index as in many cases the level of overlapping be may substantial, particularly when attempting to influence decision-makers. Hall and Miller (2010) observe that indices based on ranks often exhibit a particularly high degree of uncertainty, especially in multi-dimensional situations. They note that the data points (here, data zones) at the extreme top or bottom of an index usually show evidence of much lower uncertainty than those clustered in the middle parts of the index. Hall and Miller (2010) claim that this phenomenon is both natural and very likely to occur in many situations.

The uncertainty of ranked indices is also explored by Leckie and Goldstein (2009), for instance, with regards to school league tables. Leckie and Goldstein (2009) discovered that school rankings are typically associated with high levels of statistical uncertainty as well as there being a significant amount of overlap between schools' ranks once this uncertainty is accounted for.

Despite this, uncertainty and sensitivity analyses are valuable in that they can allow for a realistic determination of how much 'difference' between index values for various data zones is a statistically-significant difference.

## **5.2 Methods**

### **5.2.1 Monte Carlo Sampling**

There are many different ways in which to approach an uncertainty analysis for a composite index but a common method is to employ a Monte Carlo sampling algorithm. Saisana et al. (2005) and OECD (2008) both used a Monte Carlo algorithm to investigate the effects of selecting different normalisation and weighting techniques as well as quantifying the effect of removing each indicator using a one-at-a-time



approach. There are several different sampling schemes which can be implemented in a Monte Carlo analysis, listed in Helton and Davis (2000), but simple random sampling will be used in this thesis. This is due to its simplicity and its ability to produce a purely-random sample from a range of possible values. Other, more complex schemes such as stratified sampling and those based on Latin hyper-cubes (Helton and Davis, 2000) provide alternative choices but will not be considered here.

A major part of this thesis was the use of geostatistical modelling to describe the spatial patterns exhibited by the data, detailed in Chapter 3. Since all statistical models are intrinsically uncertain, a Monte Carlo sampling algorithm will be used to assess the effects of varying the modelled concentrations for each environmental determinand that comprises the EQI. The process, adapted from the methodology used by Saisana et al. (2005), involves simulating 1000 concentrations for all air, soil and water determinands for eleven data zones and then constructing 1000 index values from these simulations. The eleven data zones were the same as those shown in Figure 4.14. The index construction methodology from Chapter 4 is unchanged but the data used to create the index varies based on the standard deviations of the modelled data.

Helton and Davis (2000) observe that correlation control is very important in a sampling algorithm and that any correlation between variables being sampled simultaneously must be accounted for. The correlation structure between the individual indicators in the EQI is accounted for in the covariance matrix.

The methodology, adapted from Saisana et al. (2005), is defined below. For a random data zone:

1. For the AQI component of the index, draw  $k = 1000$  values from a multivariate normal distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$  for the four air pollutants  $\mathbf{X} = (X_1, X_2, X_3, X_4)$  i.e.  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

2. Create 1000 AQI values  $(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, x_4^{(k)})$  with appropriate normalisation, weighting and aggregation techniques, detailed in Chapter 4
3. Repeat steps 1 and 2 for SQI and WQI
4. Create EQI for  $k = 1000$  simulated values i.e.  $\text{EQI}^{(k)}$  from  $\text{AQI}^{(k)}$ ,  $\text{SQI}^{(k)}$  and  $\text{WQI}^{(k)}$

This algorithm is run for all three indices (the EQI created using the ranking, re-scaling and z-score normalisation methods) as well as equivalent indices constructed using equal weights for all indicators and domains.

The macroinvertebrate data are not included in this process and all 1000 indices will contain the modelled macroinvertebrate probability from Chapter 3. This is due to added complexity in the distribution of its values and possible difficulties in interpretation.

This method attempts to quantify the effects of varying the concentration for each determinand in the index for each of the chosen eleven data zones. It also assumes that the position of all other data zones in the index remains fixed, which may be an unrealistic assumption in practice, particularly with regards to air quality. It is likely that if a local area is affected by a sudden change in air quality that the data zones surrounding it will also be affected.

### 5.2.2 Kappa Statistics

One way in which environmental quality indices can be compared from year-to-year is by using Cohen's kappa (Cohen, 1960), an inter-rater reliability measure that aims to quantify the level of 'agreement' between two 'observers', which may be human observers or, in this case, different index construction methodologies. It is closely related to intra-class correlation measures. Cohen's kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \quad (5.1)$$

where  $p_o$  is the observed agreement amongst the raters and  $p_e$  is the probability of chance agreement calculated using the observed data. If the raters agree completely, then  $\kappa = 1$ ; if there is no agreement whatsoever, then  $\kappa = 0$ .

The weighted kappa statistic (Cohen, 1968; Fleiss et al., 1969), which may be interpreted as chance-corrected weighted proportional agreement (Steltner et al., 2002), is more useful when the categories are ordinal (i.e. data zone ranks from 1 to 1,748). It alters the statistic produced by Equation 5.1 by assigning less weight to agreement as categories are further and further apart (Vierra and Garrett, 2005). There is no one weighting scheme used in the literature but the R package **psych** (Revelle, 2015) used in this thesis assumes quadratic weights, a common choice. The formula for the weighted kappa is defined in Equation 5.2.

$$\kappa = 1 - \frac{\sum_{i=1}^k \sum_{j=1}^k w_{ij} x_{ij}}{\sum_{i=1}^k \sum_{j=1}^k w_{ij} m_{ij}} \quad (5.2)$$

where  $w_{ij}$ ,  $x_{ij}$  and  $m_{ij}$  are the elements of the weight, observed and expected matrices, respectively. Fleiss et al. (1969) provided a corrected weighted kappa statistic, which can also be used to construct confidence intervals, and is the result produced by Equation 5.2.

There is no single way in which to interpret a kappa statistic but several rules of thumb have been developed with one of the most commonly-used approaches being summarised by Vierra and Garrett (2005), shown in Table 5.1.

Table 5.1: Categories for interpreting Cohen’s kappa statistic (Vierra and Garrett, 2005)

<b>Kappa</b>	<b>Agreement</b>
< 0.01	Less than chance agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1	Almost perfect agreement

## 5.3 Results

### 5.3.1 Uncertainty Analysis

The Monte Carlo-based uncertainty analysis described in Section 5.2.1 was implemented for all three indices constructed in Chapter 4, i.e. using the

- Ranking normalisation method (Figure 4.11)
- Re-scaling normalisation method (Figure 4.12)
- Z-score normalisation method (Figure 4.13)

An equivalent index was created for all three normalisation methods using equal weights for all indicator and domains, since the weighting scheme for a composite index is an important step in its construction.

Figures 5.1, 5.2 and 5.3 show the results of the uncertainty analyses for the eleven selected data zones, restricted to the 5<sup>th</sup> and 95<sup>th</sup> percentiles of the distributions.

Figure 5.1(a) shows the results for the index constructed using the ranking normalisation method with factor analysis weights. As mentioned in Section 5.1, the data zones at the lower and upper extremities in relation to environmental quality typically exhibit much lower levels of uncertainty when compared to data zones in

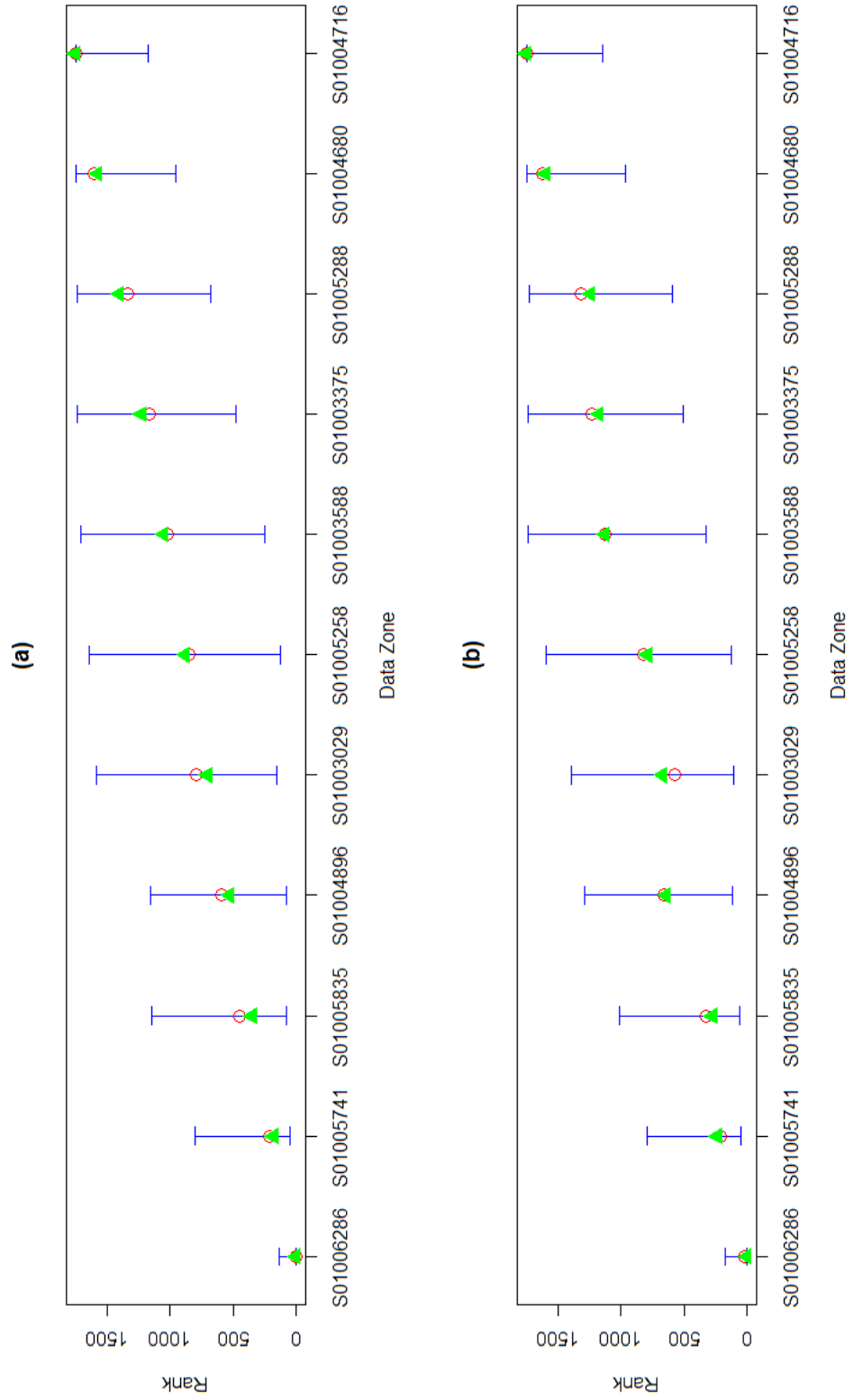


Figure 5.1:  $5^{th}$  and  $95^{th}$  percentile intervals for ranking normalisation method:  
(a) factor analysis weights; (b) equal weights (Legend:  $\blacktriangle$  = EQI rank;  $\circ$  = median)

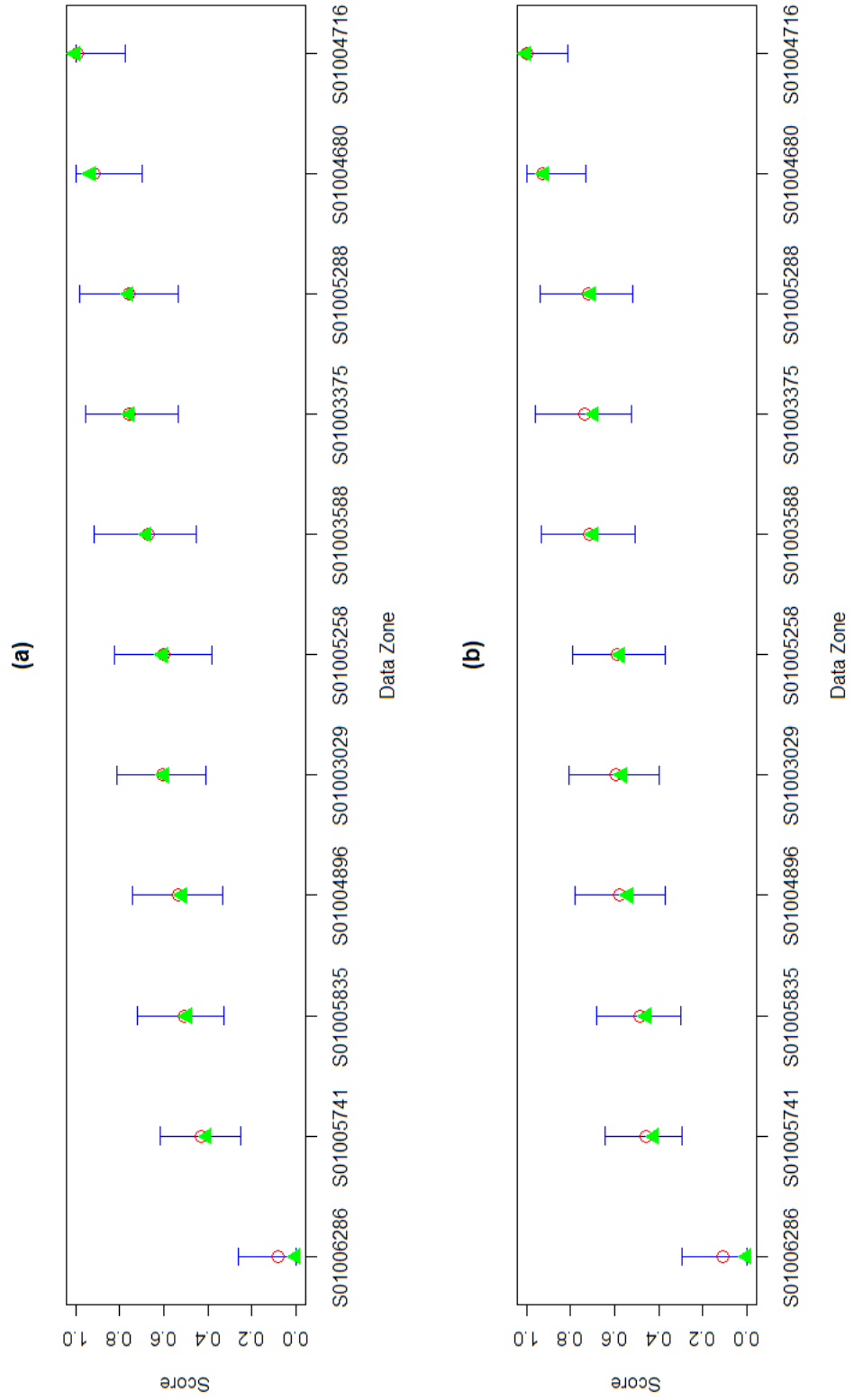


Figure 5.2:  $5^{th}$  and  $95^{th}$  percentile intervals for re-scaling normalisation method:  
(a) factor analysis weights; (b) equal weights (Legend:  $\blacktriangle$  = EQI rank;  $\circ$  = median)

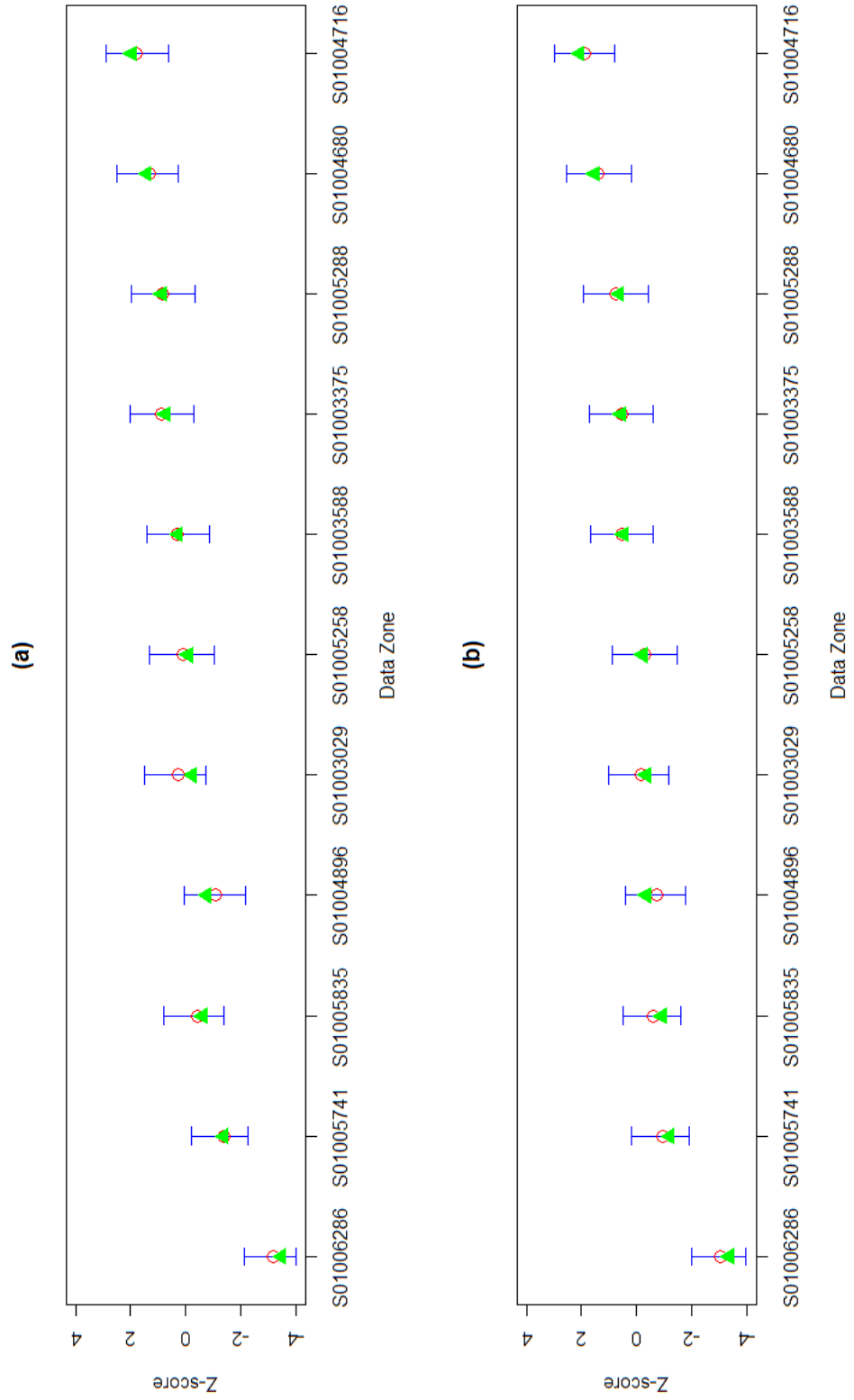


Figure 5.3: 5<sup>th</sup> and 95<sup>th</sup> percentile intervals for z-score normalisation method:  
(a) factor analysis weights; (b) equal weights (Legend:  $\blacktriangle$  = EQI rank;  $\circ$  = median)

the middle of the index, for instance data zone S01005258. The uncertainty analysis also shows that the index values constructed in Chapter 4 are typically situated close to the median of all calculated ranks from the Monte Carlo analysis. Figure 5.1(b) shows somewhat similar results to Figure 5.1(a), indicating that using either equal weights or weights drawn from factor analysis for the individual indicators may not necessarily affect the subsequent index values. However, some data zones exhibit lower uncertainty when assuming equal weights whilst others show evidence of higher uncertainty.

The uncertainty analysis for the z-scores normalisation method, shown in Figure 5.2(a), appears to show that the data zone index values exhibit much lower uncertainty using this method, implying that the z-score method may be more trustworthy. Like the ranking method, the EQI index values are largely similar to the median of the Monte Carlo samples. Again, Figure 5.2(b) appears to show that assuming equal weights for all indicators does not cause much change in the final results.

Figure 5.3(a) appears to show evidence of less uncertainty for the re-scaling method, overall, when compared to the ranking method but is less successful than the z-score approach. Figure 5.3(b) again implies that equal weighting does not unduly affect the results of the index.

### **5.3.2 Comparing Indices Over Time**

It is of interest to many, including policymakers, to assess whether environmental quality is improving or deteriorating over time. Composite indices can be used for such quantitative assessments. This section will look at how the various environmental quality indices constructed in Chapter 4 vary over a five-year period, from 2008 to 2012. This will comprise time series plots (with an associated estimate of uncertainty based on the results presented in Section 5.3.1) and Cohen's weighted kappa statistics.



As aforementioned, the soil quality data were only available for one year and were assumed to be largely constant over time for the purposes of this thesis.

### **Time Series Plots**

The indices constructed in Chapter 4 were for the year 2011 only. The indices were then re-created for the years 2008, 2009, 2010 and 2012. Time series plots were constructed for three data zones out of the eleven that Section 5.2 also focussed on: S01003029, S01004896 and S01006286. Uncertainty estimates for the EQI were drawn from the Monte Carlo samples constructed in Section 5.3.1.

Figures 5.4(a), 5.4(c) and 5.4(e) show how the EQI varies over the period in question for all three normalisation methods (ranking, re-scaling and z-scores, respectively) for the data zone S01003029. A similar pattern can be observed with environmental quality appearing to worsen very gradually from 2008 to 2011 before improving in 2012. The re-scaling method in particular highlights a sudden decline in environmental quality for the year 2011 when compared with 2010 but this effect is not so pronounced for the other two methods. The 5<sup>th</sup> and 95<sup>th</sup> percentiles (red and blue, respectively) for each year also highlight the differences in the levels of uncertainty exhibited by each of the three normalisation methods, discussed in Section 5.2. The ranking method (Figure 5.4(a)) exhibits, by far, the greatest level of uncertainty, shown by the very wide intervals. The re-scaling method (Figure 5.4(c)) shows evidence of less uncertainty and the z-score method (Figure 5.4(e)) shows even less uncertainty.

Figures 5.4(b), 5.4(d) and 5.4(f) are replicates of Figures 5.4(a), 5.4(c) and 5.4(e), respectively, except with the year 2008 being changed to the value zero. This is to gain an impression as to how much change there is from year-to-year compared to a baseline of zero. Again, there is some evidence that environmental quality may remain at the same level until 2012 when there is a decline, but this is not shown by the z-score method (Figure 5.4(f)).

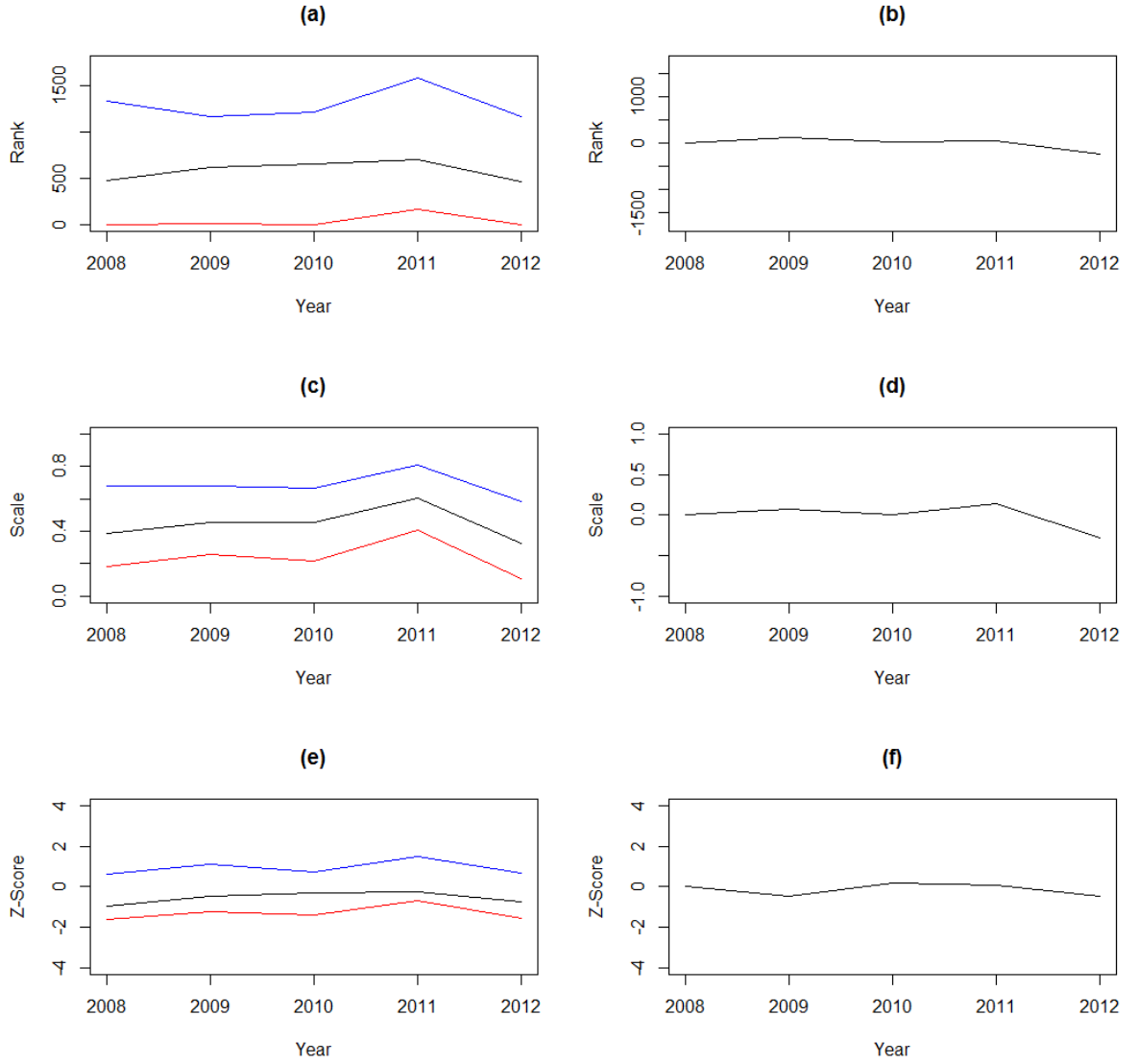


Figure 5.4: Time series plots (2008 – 2012) for DZ S01003029: (a) EQI for ranking method with uncertainty intervals; (b) Ranked EQI change over time (2008 = 0); (c) EQI for re-scaling method with uncertainty intervals; (d) Re-scaled EQI change over time (2008 = 0); (e) EQI for z-score method with uncertainty intervals; (e) EQI for re-scaling method with uncertainty intervals; (f) Z-score EQI change over time (2008 = 0) (Legend: — = 95<sup>th</sup> percentile; — = 5<sup>th</sup> percentile)

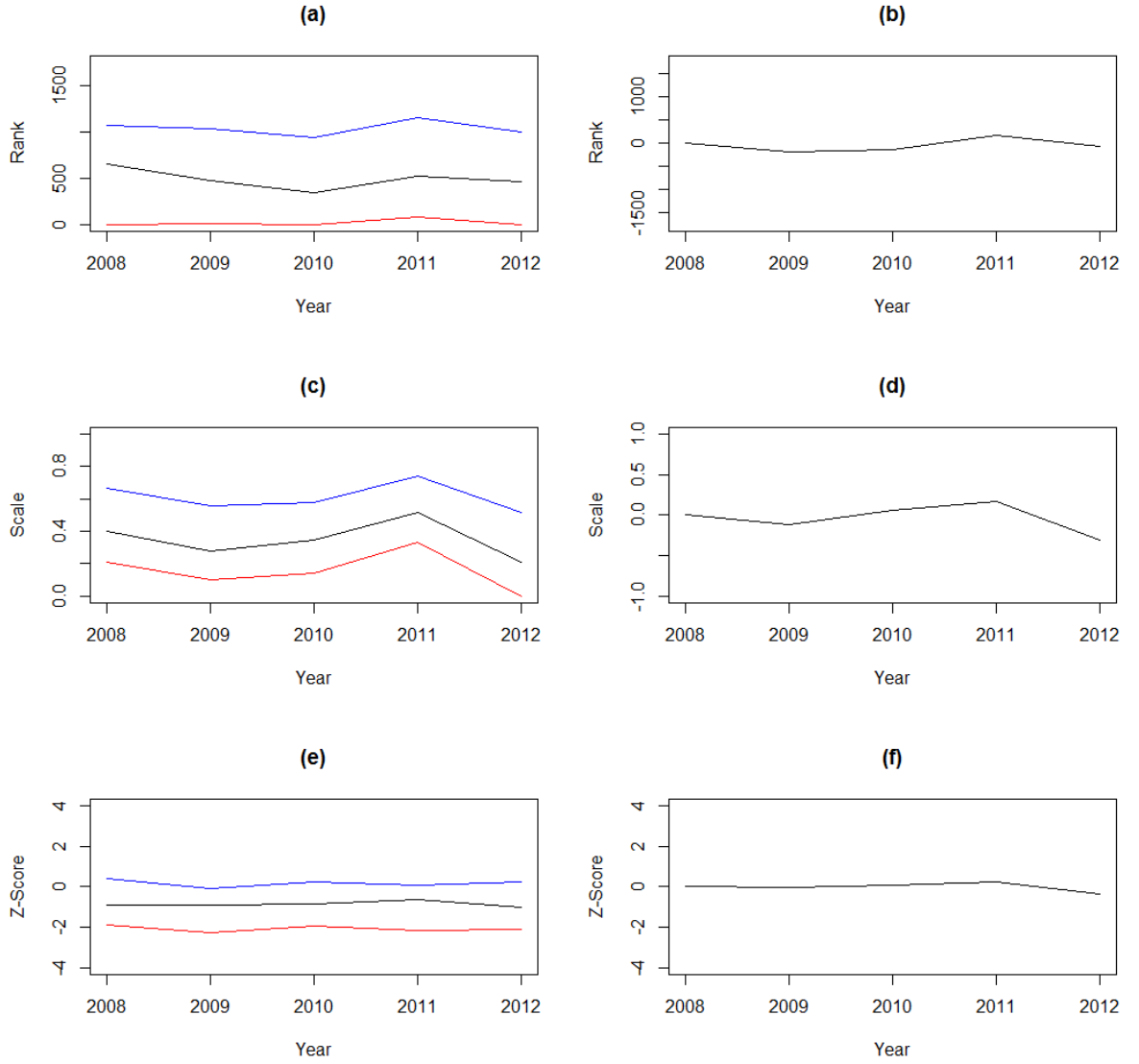


Figure 5.5: Time series plots (2008 – 2012) for DZ S01004896: (a) EQI for ranking method with uncertainty intervals; (b) Ranked EQI change over time (2008 = 0); (c) EQI for re-scaling method with uncertainty intervals; (d) Re-scaled EQI change over time (2008 = 0); (e) EQI for z-score method with uncertainty intervals; (f) EQI for re-scaling method with uncertainty intervals; (f) Z-score EQI change over time (2008 = 0) (Legend: — = 95<sup>th</sup> percentile; — = 5<sup>th</sup> percentile)

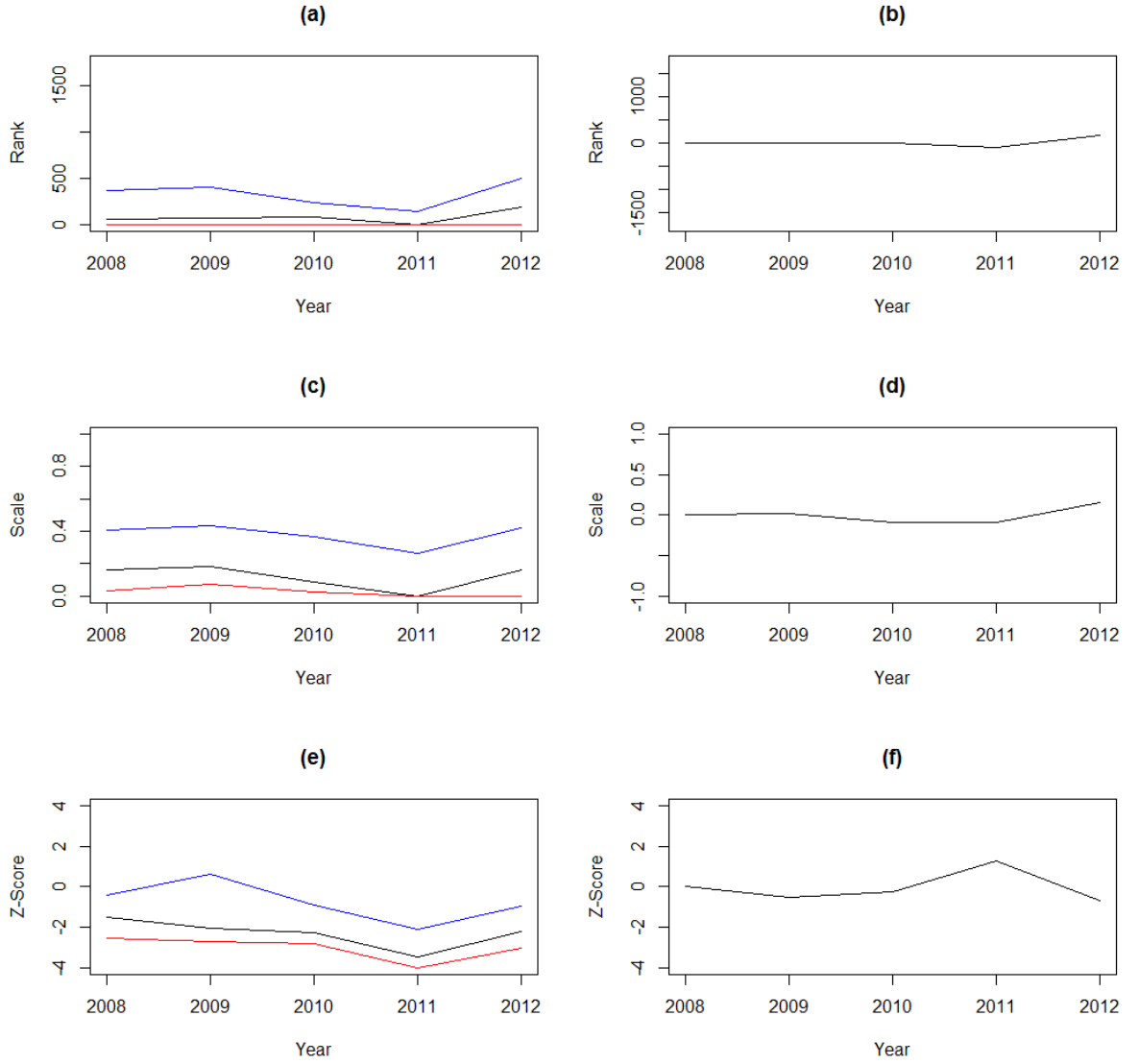


Figure 5.6: Time series plots (2008 – 2012) for DZ S01006286: (a) EQI for ranking method with uncertainty intervals; (b) Ranked EQI change over time (2008 = 0); (c) EQI for re-scaling method with uncertainty intervals; (d) Re-scaled EQI change over time (2008 = 0); (e) EQI for z-score method with uncertainty intervals; (f) EQI for re-scaling method with uncertainty intervals; (f) Z-score EQI change over time (2008 = 0) (Legend: — = 95<sup>th</sup> percentile; — = 5<sup>th</sup> percentile)

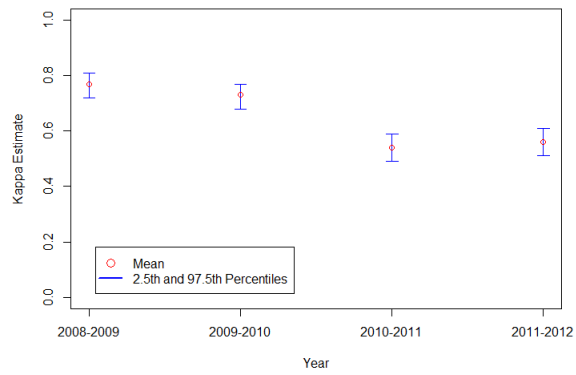
Figure 5 shows a similar set of plots for data zone S01004896 although there is conflicting views as to how environmental quality changes from 2008 to 2012 based on which normalisation method is used. The ranking method (Figure 5.5(a)) shows evidence of improvement in environmental quality until 2010 when it begins to worsen again. However, the re-scaling method (Figure 5.5(c)) suggests that environmental quality generally worsens until 2011 before improving in 2012. The z-score method (Figure 5.5(e)) shows little evidence of change. These patterns are broadly confirmed by the respective plots showing change over time (Figures 5.5(b), 5.5(d) and 5.5(f)).

For data zone S01006286, the three methods generally agree, with Figures 5.6(a), 5.6(c) and 5.6(e) suggesting that environmental quality gradually improves before declining again in 2012. Since data zone S01006286 is at the lower end of the three indices, the uncertainty intervals are much narrower than those for data zones S01003029 and S01004896, as discussed in Section 5.3.1. The uncertainty intervals are particularly narrow for the index constructed using the ranking method. Again, the temporal changes shown in Figures 5.6(b), 5.6(d) and 5.6(f) largely agree with the prior inference that environmental quality gradually improves until 2012.

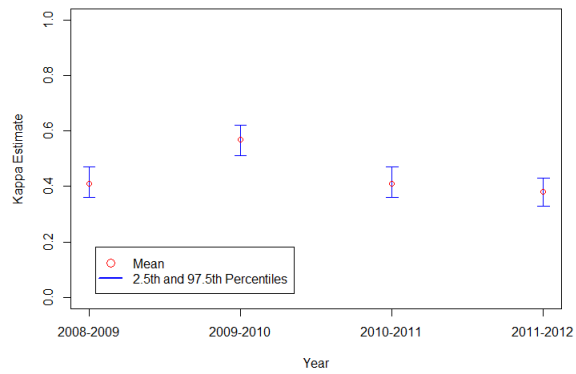
### **Kappa Statistics**

Cohen’s weighted kappa statistic was also used to observe how environmental quality varies between 2008 and 2012, using the level of agreement from year to year as an indication of change. Plots were constructed for all three normalisation techniques. It should be noted that the results from the re-scaled and z-scores methods were also ranked from 1 to 1,748 in order for the kappa statistic to be used. How the 1,748 data zones are ordered is of interest here, so this step should be reasonable.

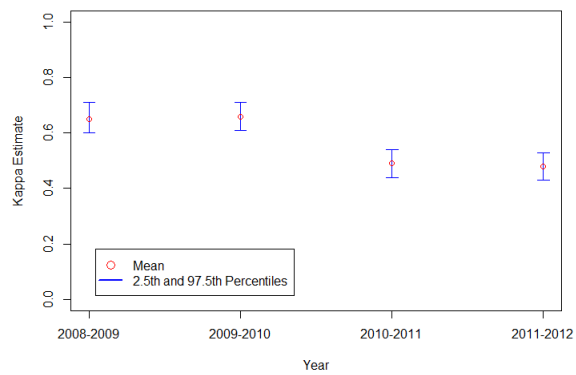
The EQI, shown in Figure 5.7, generally shows evidence of declining agreement from 2008–2009 to 2011–2012, indicating change in data zone ranks over this period. However, inspection of the uncertainty intervals (constrained to the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles) implies that these changes may not necessarily be statistically-significant



(a) Ranking



(b) Re-scaling



(c) Z-scores

Figure 5.7: Plots showing levels of agreement for EQI from year-to-year (2008 – 2012)

with regards to all three normalisation methods. For instance, in Figure 5.7(b), the uncertainty bands generally overlap each other from year-to-year, suggesting that changes in the results for the re-scaling method may not be statistically-significant.

The ranking and z-scores methods show evidence of much higher agreement overall than the re-scaling approach with higher kappa values being observed. This means that there are perhaps less pronounced changes in environmental quality from year to year for these two methods when compared to the re-scaling approach. For both the ranked and z-score indices, there appears to be a statistically-significant change from 2009–2010 to 2010–2011, as the uncertainty intervals do not overlap. This may suggest that data zone ranks have moved over this period.

Interpreting these results with regards to the categories shown in Table 5.1, for Figures 5.7(a) and 5.7(c), there is typically ‘substantial agreement’ from 2008 to 2010 but ‘moderate agreement’ from year-to-year afterwards. This may be an indication of change in data zone positions within the index over the entire time period. There is evidence of greater change in Figure 5.7(b) since there is generally only ‘fair’ to ‘moderate’ agreement from one year to the next.

## 5.4 Chapter Summary

This chapter has employed a variety of methods found in the literature for assessing composite indices, focussing on quantifying the level of statistical uncertainty within the indices and observing how each index varies over time. The purpose of this was to assess the ‘robustness’ of the index and to find ways of checking whether any year-to-year changes in the index were statistically-significant.

To quantify uncertainty within the index, an approach derived from the family of ‘uncertainty and sensitivity analyses’ has been used. The uncertainty analysis focussed on eleven data zones, as an illustration. Monte Carlo sampling was utilised to

construct a distribution of 1000 possible values for each indicator within the index. This led to 1000 possible index values which was then restricted to the 5<sup>th</sup> and 95<sup>th</sup> percentiles so that an interval of uncertainty was calculated for the index values of the eleven selected data zones.

As indicated by literature such as Hall and Miller (2010), uncertainty intervals were typically narrower for data zones at the extreme upper and lower tails of the index. For the ranking method especially, the intervals were often significantly wide for data zones in the middle of the index.

For the purposes of checking how the environmental quality indices varied over time, a combination of time series plots and Cohen's weighted kappa statistics was used. The analysis focussed on a five-year period from 2008 to 2012 and, for the time series plots, three data zones out of the eleven from the uncertainty analysis were considered.

For the time series plots, the intervals from the uncertainty analysis were also used so that it could be determined how much change in environmental quality from year-to-year was statistically-significant. Changes in environmental quality were dependent on the data zone in question and no common pattern was observed between the three data zones.

Cohen's weighted kappa statistics were used to assess the level of statistical 'agreement' from year-to-year in terms of the environmental quality indices. The three different normalisation methods typically showed similar patterns with regards to annual changes.

The purpose of using time series plots and Cohen's weighted kappa was ultimately to detect any change in the indices over time and whether any year-to-year changes were statistically-significant or due to random variation. In a broader sense, employing



all of the above methods to critique the indices constructed in Chapter 4 serves to enhance their practical use and allow for better inference once uncertainty is accounted for.

# Chapter 6

## Discussion and Conclusions

The primary objective of this thesis was to investigate and develop methods for constructing a multidimensional, composite environmental quality index at a small spatial scale. Specifically, the index was created for the purposes of measuring environmental quality in Greater Glasgow at data zone resolution. The index was constructed at an annual level and incorporated 1,748 data zones across seven local authorities in the Greater Glasgow region and was created for the year 2011 only, at first.

The index construction methodology described in this thesis followed a set of guidelines issued by the OECD (2008), who provided ten key steps that should be considered when creating a composite index. Important stages include the selection of indicators, data quality assessments and choosing appropriate normalisation, weighting and aggregation schemes for constructing the index, as well as attempting to quantify any uncertainty in the results.

### **Theoretical Framework**

The index was comprised of three individual domains (or sub-indices), each measuring a different aspect of the environment: air, soil and water. Each domain was constructed from a number of individual indicators for common air, soil and water determinands. From these indicators and domains, an impression of the general

environmental quality in a particular area can be gained with the best-performing areas typically performing well across most, if not all, indicators.

### **Data Selection and Quality**

The air quality component of the index (referred to as the AQI or ‘air quality index’) consisted of four pollutants:  $NO_2$ ,  $NO_X$ ,  $PM_{10}$  and  $PM_{2.5}$ . Data to measure these four determinands were collected from publicly-available sources, including websites operated by the Scottish Government and the Department for Environment, Food and Rural Affairs.  $NO_2$  diffusion tube data collected by local authorities were also used.

The soil quality domain (the SQI) consisted of five determinands: arsenic, chromium, lead, nickel and selenium. The data were provided by the British Geological Survey (Fordyce et al., 2012) but were only available for one year.

The water quality domain (the WQI) included information on six variables: ammonium, copper, dissolved oxygen, soluble reactive phosphorus, pH and macroinvertebrate activity (measured as the average score per taxon). The data were much more spatially-sparse than those for air and soil quality and then necessitated the use of statistical modelling, discussed below.

### **Exploratory Analysis**

Firstly, an in-depth spatial and temporal exploratory analysis was conducted on all data in order to gain an overall impression as to how the environmental processes varied over space and time. Generally, the poorest environmental quality (across all three domains) was observed in the highly-urbanised data zones of central Glasgow and central Lanarkshire (particularly in areas such as Motherwell and Hamilton) with better environmental quality typically being found in rural areas. However, rural areas to the west of Glasgow also appeared to suffer from poorer soil quality when compared to rural areas to the north and south.

There was not much visual evidence of any significant, long-term changes in any of the air or water pollutant concentrations in a temporal sense. Again, the soil data were only available for one year and were assumed to have not seen much year-to-year change in this or any subsequent temporal analyses.

### **Modelling of Data**

Since the data were not available at a spatial resolution high enough to allow for an assessment of environmental quality in all 1,748 data zones, the data were then modelled to create spatial surfaces. All air, soil and water quality determinands (except for the macroinvertebrate data) were modelled using isotropic smooths with an appropriately-selected smoothing parameter and restricted maximum likelihood for parameter estimation. The model fits were assessed using a variety of plots based on the model residuals, including whether the residuals followed a Gaussian distribution.

The macroinvertebrate data were modelled using a random forest process since they were only available on a categorical scale. The data were reduced to a binary classification (‘good’ or ‘bad’) by merging the five individual categories into two in order to allow for more balanced categories.

The models were then used to predict data at every point in a specified rectangular grid which covered the study region so that each data zone would have at least one ‘observation’. However, some very geographically-small data zones in central Glasgow required a prediction location to be manually-specified since they were missed by the prediction grid. A single value for each data zone was then calculated by taking the arithmetic mean of all predictions.

### **Normalisation, Weighting and Aggregation of Indicators**

The actual construction of composite indices usually involves three stages: normalisation (so that all indicators are on the same mathematical scale), weighting (in

order to avoid multicollinearity issues or if some indicators are more important than others) and aggregation (how the indicators are combined mathematically).

The modelled data were used to construct an environmental quality index for each data zone by first creating the three individual sub-indices for air, soil and water. Three different normalisation methods were considered: z-scores, re-scaling to  $[0, 1]$  and ranking, the latter following the approach used by the Scottish Index of Multiple Deprivation (Scottish Government, 2012).

Weights for each indicator and domain were then calculated using factor analysis although the macroinvertebrate indicator was given an arbitrary weighting. Finally, an arithmetic aggregation method was employed to create a final value for each data zone for air, soil and water. This process was repeated in order to amalgamate the three domains into the final environmental quality index (EQI).

## **Main Results**

The results for the air quality index (AQI) appeared to show, as one would expect, that the poorest air quality was found in highly-urbanised areas such as the centre of Glasgow and in other major towns. The soil quality index (SQI) suggested that the data zones with the highest level of pollution were typically found, again, in central Glasgow but also in the far-west of the study region in the local authority of Renfrewshire. Finally, the water quality (WQI) implied that the poorest water quality was observed in the River Clyde watercourse, especially in the major towns to the east of Glasgow and within the city itself.

Visualising the results of the EQI appeared to confirm that poorer environmental quality was typically found in urban areas with better environmental quality being observed in rural areas. This was a consistent phenomenon across all three scales (ranks, z-scores and the  $[0, 1]$  scale).

### **Uncertainty Analysis**

The indices were then subject to an uncertainty analysis in order to quantify any uncertainty in the index results. This was achieved using a Monte Carlo sampling algorithm on eleven selected data zones by using the model standard errors to generate a plausible range of possible values for each determinand. Intra-determinand correlation was also considered during the simulations. It was discovered, as is typically found in the literature, that the index based on ranks suffered from quite high levels of uncertainty, particularly in the data zones towards the middle of the index. Data zones at the lower and upper tails of the distribution, however, appeared to exhibit lower uncertainty, as is also frequently found in the literature. The z-score and re-scaling methods showed good evidence of much lower uncertainty but again those data zones in the middle of the indices were the most uncertain.

### **Comparing Environmental Quality Over Time**

Finally, the index was reconstructed for the years 2008, 2009, 2010 and 2012 in order to quantify how environmental quality changes over a period of five years, acknowledging that the soil quality data were only available for one year. This was achieved through time series plots (including uncertainty intervals drawn from the results of the Monte Carlo simulations described above) and using Cohen's weighted kappa statistics. Very little change was observed in the environmental quality index results of the eleven selected data zones since all index values fell within the previous and next year's uncertainty intervals, meaning there was little evidence of significant change from year-to-year.

## **6.1 Further Work**

There remains scope for significant further work in this area, particularly as the field of composite indices has become an increasingly popular area of research in recent years. Also, there remain a few technical details in this thesis which would benefit from further attention.

In Section 3.5.1, Euclidean distance was used to assess the presence, or absence, of significant spatial correlation in the model residuals. However, several papers have argued that Euclidean distance is not the most effective way to assess correlation in river networks. O'Donnell et al. (2014), for instance, propose alternative methods but were not considered here due to time constraints.

Another possible extension would be to consider other normalisation methods when constructing the index. Nardo et al. (2005) and OECD (2008) list several other possibilities, including measuring each determinand in relation to an acceptability threshold set by policymakers. This would allow one to see which pollutants are in violation of air, soil or water quality guidelines and in which data zones. Key examples of this sort of index scale are the water and soil indices created by the Canadian Council of Ministers of the Environment (CCME, 2001; CCME, 2007), which were discussed in Chapter 1. Also, weighting schemes other than factor analysis, including those determined by consultation with experts, could also be considered and are listed in OECD (2008).

In Chapter 3 of this thesis, after the data were modelled and predictions were made, the arithmetic mean was used to calculate a predicted value for each data zone. However, other choices such as the median or the maximum may be appropriate in other situations. The median may provide a more statistically-robust value for each data zone if there are significant outliers in the predicted data. The maximum may be appropriate if one wished to report on a 'worst-case-scenario' in each data zone.

For spatial data analysis, there is a need for data of high spatial resolution across all determinands, which the water quality component of the indices developed in this thesis somewhat lacked, especially when compared to the air and soil data sets. Having such data would mean that predictions are more likely to accurately reflect the observed spatial processes across the study region and be subject to less statistical error. Similarly, the soil quality data used for the analyses presented

in this thesis were only available for one year, which may have had an impact on some results and inference when the indices were being compared from year to year. Ascertaining regularly-measured soil data is difficult but having such data would obviously be an important step forward for a project such as this.

The time series plots for environmental quality displayed in Section 5.3.1 could also be an avenue for further development, particularly in relation to the intervals of uncertainty for each data zone in each year. The intervals shown in this thesis were drawn from the Monte Carlo-based uncertainty analysis in Chapter 5, which may not necessarily be the most effective approach. Including a formal sensitivity analysis alongside the uncertainty analysis shown in Chapter 5, detailed in literature such as Saltelli et al. (2000), is a very obvious field for further work. Also, as mentioned in Section 5.2.1, the uncertainty analysis assumed that the position of all data zones in the index would be fixed apart from the one being considered, whose pollutant concentrations were allowed to vary. This assumption is probably unrealistic as environmental conditions in one area may affect other surrounding areas. Including this possibility as a factor in the uncertainty analysis may be an opportunity for further development.

Finally, the index construction framework outlined in Table 1.2 included a step which involved comparing an index to another similar example. A major component of Chapter 4 of this thesis was the construction of an environmental quality index based on the SIMD (Scottish Government, 2012) method. It may be possible to informally compare the results of these two indices, focussing on the 1,748 data zones that this thesis concentrated on. It is important to note that the Scottish Index of Multiple Deprivation (2012) is not an environment-focussed index and, hence, comparing it with the EQI is simply for informative purposes only. The results for the EQI (for the year 2012 and based on the SIMD method) and the SIMD results for 2012 are displayed below in Figure 6.1. It should be noted that the SIMD ranks have been inverted in Figure 6.1(b), since rank 1 in SIMD indicates the



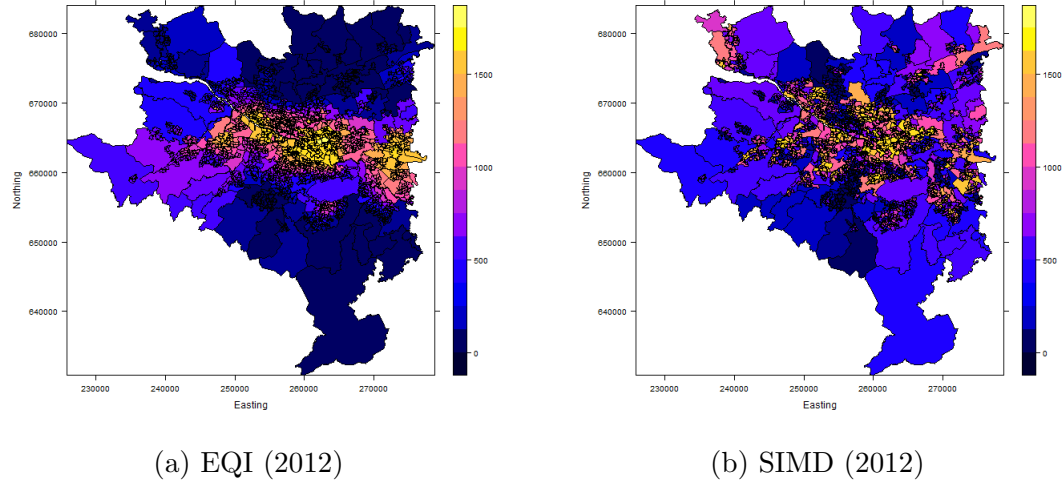


Figure 6.1: Comparing the EQI and the Scottish Index of Multiple Deprivation (Scottish Government, 2012) – contains Ordnance Survey data © Crown copyright and database right (2014)

worst-performing data zone, whilst in the EQI rank 1 indicates the best-performing data zone.

From the informal comparison of the indices in Figure 6.1, it can be observed that the patterns are largely different with the areas of worst deprivation in Figure 6.1(b) not necessarily exhibiting the poorest environmental quality in Figure 6.1(a). Nevertheless, investigating any possible links between environmental quality and deprivation would be an interesting avenue to explore, perhaps using methods for testing spatial homogeneity between two spatial patterns.

All of the potential improvements or alternative choices for index construction discussed above would be appropriate to consider. As aforementioned, the field of composite indices is a very active area of study and other possibilities to improve or change the indices presented in this thesis can be found in the ever-expanding literature.

# Appendices

## A. Additional Material Related to Chapter 2

Table A1: Descriptive statistics for selected  $NO_X$  air quality monitoring stations in 2011 (units of measurement =  $\mu\text{g}/\text{m}^3$ )

Monitoring Station	Mean	St. Dev.	Min	Q1	Median	Q3	Max
Glasgow Byres Road	43.55	45.19	0.00	18.00	33.00	54.00	190.00
Glasgow Centre	16.42	30.99	0.00	3.00	6.00	16.00	466.00
Glasgow Waulkmillglen Reservoir	3.95	12.29	0.00	1.30	1.30	2.50	225.00
Bearsden	42.61	49.96	0.00	11.00	29.00	55.00	531.00
Clydebank	16.13	29.41	0.00	1.00	4.00	18.00	365.00
East Kilbride	35.75	45.18	0.00	10.00	23.00	43.00	670.00

Table A2: Descriptive statistics for all  $PM_{2.5}$  air quality monitoring stations in 2011 (units of measurement =  $\mu\text{g}/\text{m}^3$ )

Monitoring Station	Mean	St. Dev.	Min	Q1	Median	Q3	Max
Glasgow Centre	10.17	10.94	0.00	5.00	8.00	12.00	288.00
Glasgow Kerbside	22.39	14.12	0.00	13.00	19.00	28.00	286.00

Table A3: Descriptive statistics for selected  $PM_{10}$  air quality monitoring stations in 2011 (units of measurement =  $\mu\text{g}/\text{m}^3$ )

Monitoring Station	Mean	St. Dev.	Min	Q1	Median	Q3	Max
Glasgow Byres Road	23.46	18.60	0.00	13.00	19.00	29.00	249.00
Glasgow Centre	16.60	12.24	0.00	10.00	14.00	19.00	316.00
Glasgow Waulkmillglen Reservoir	12.11	7.62	0.00	7.30	10.00	15.00	62.00
Bearsden	19.90	18.50	0.00	8.00	16.00	27.00	352.00
Clydebank	16.92	11.53	0.00	9.00	14.00	21.00	115.00
East Kilbride	15.53	11.96	0.00	8.00	13.00	20.00	183.00

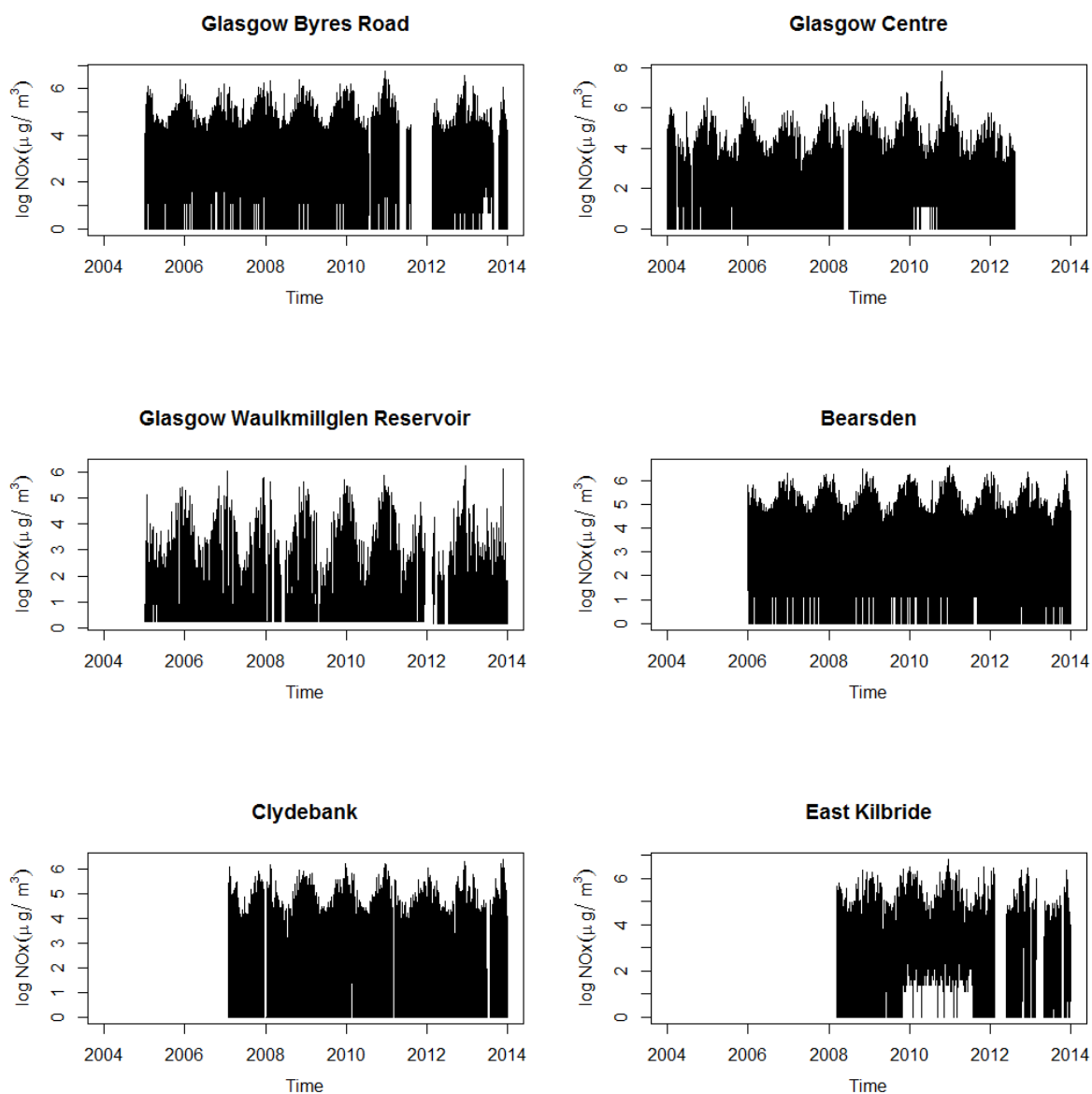


Figure A1: Time series plots of log-transformed  $NO_x$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

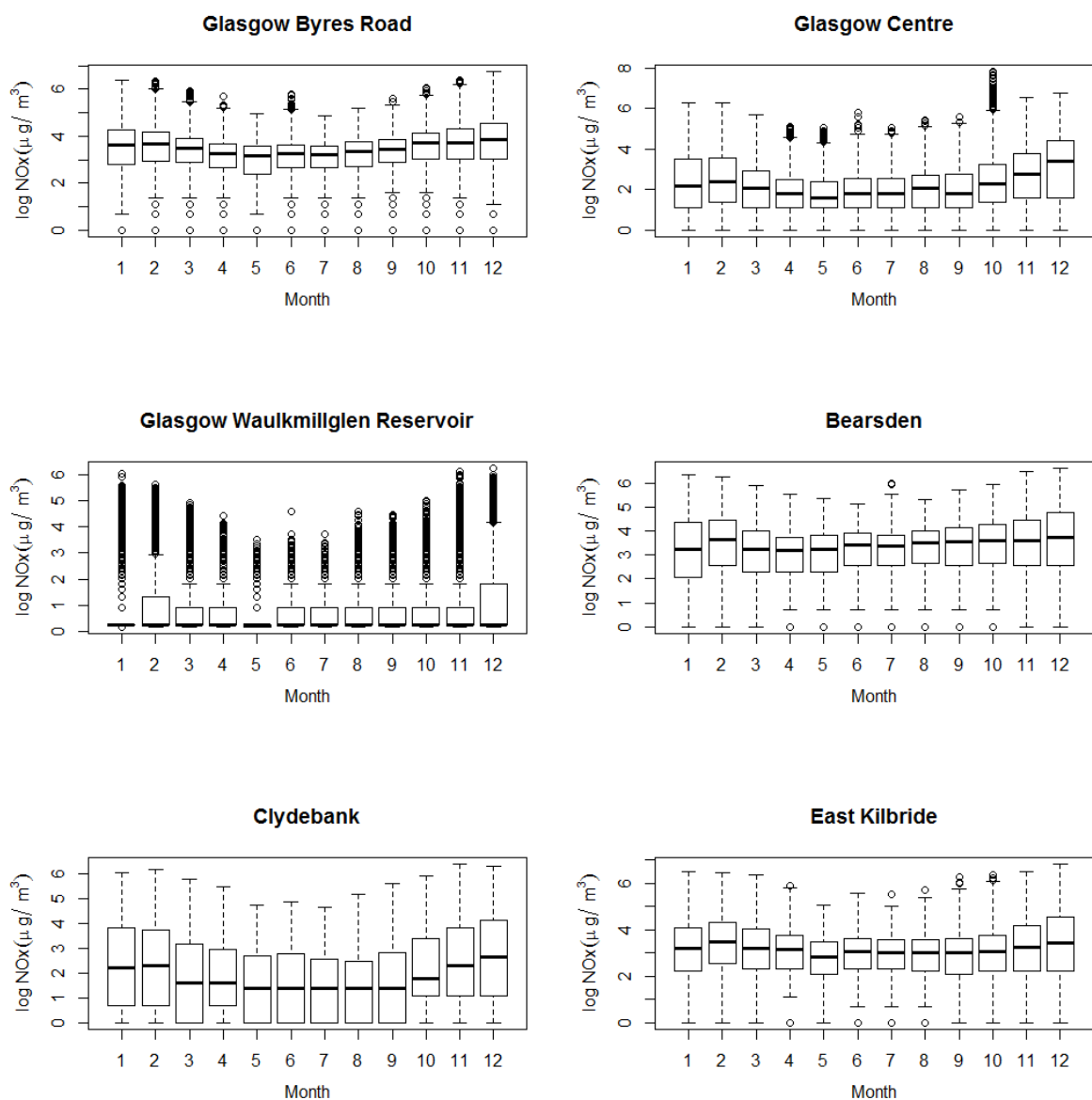


Figure A2: Boxplots of monthly log-transformed  $NO_x$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

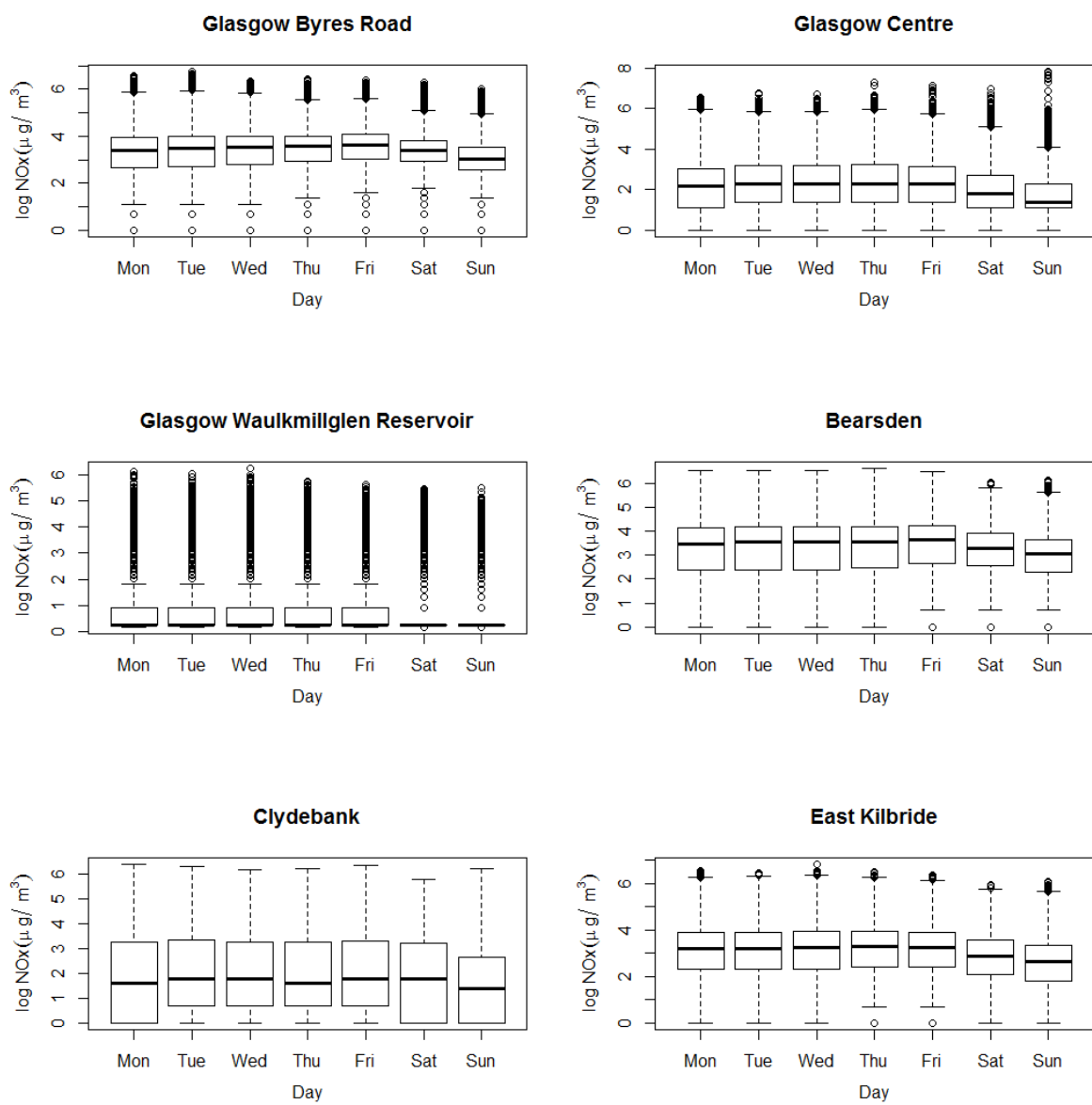


Figure A3: Boxplots of day-within-week log-transformed  $\text{NO}_x$  concentrations ( $\mu\text{g}/\text{m}^3$ ) at selected monitoring stations

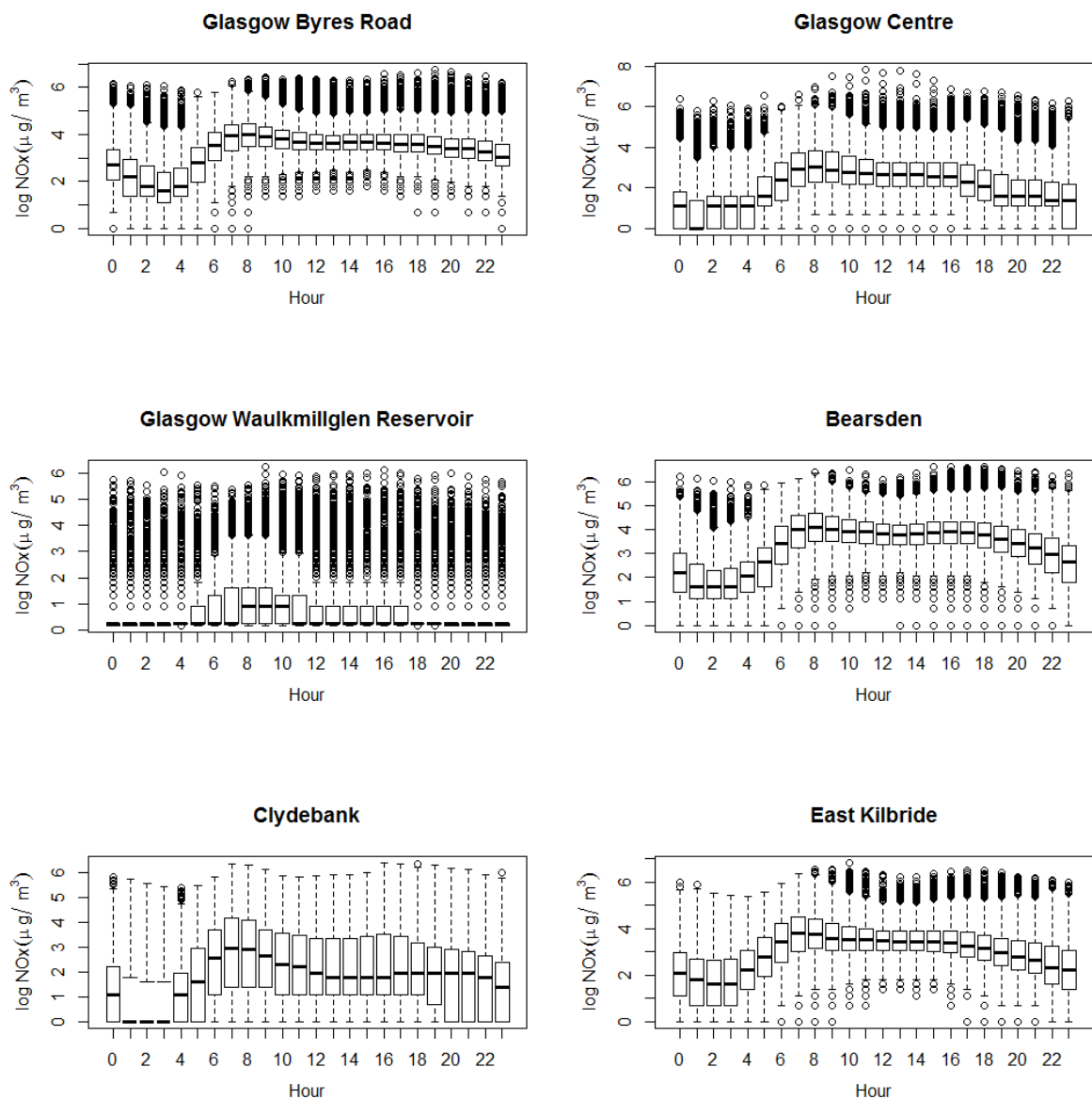


Figure A4: Boxplots of hourly log-transformed  $NO_X$  concentrations ( $\mu\text{g}/\text{m}^3$ ) at selected monitoring stations

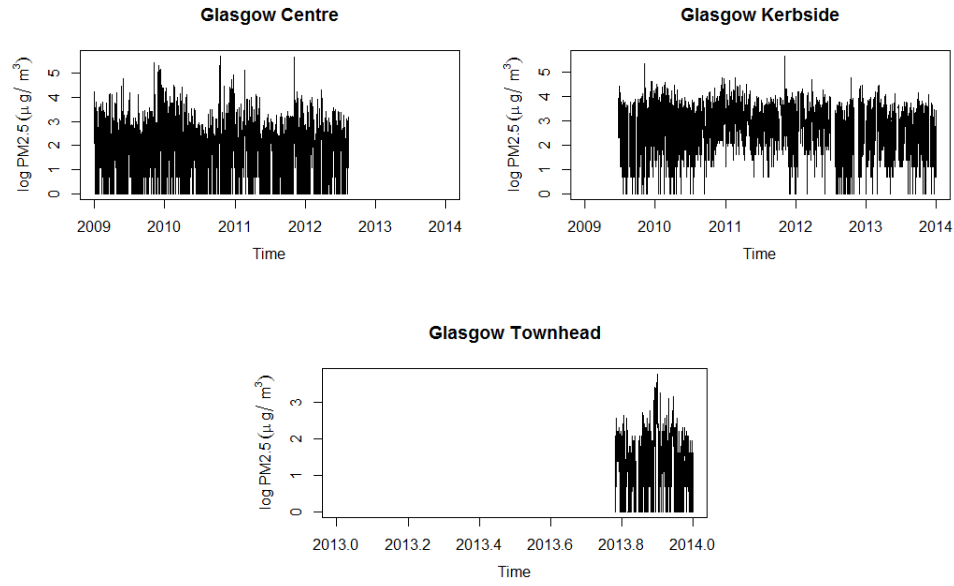


Figure A5: Time series plots of log-transformed  $PM_{2.5}$  concentrations ( $\mu\text{g}/\text{m}^3$ ) at all monitoring stations

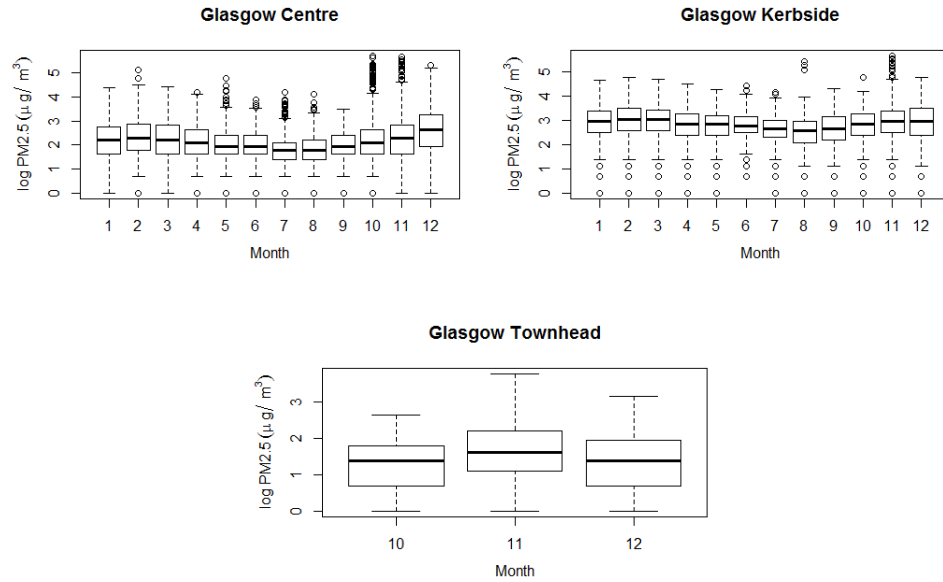


Figure A6: Boxplots of monthly log-transformed  $PM_{2.5}$  concentrations ( $\mu\text{g}/\text{m}^3$ ) across all monitoring stations

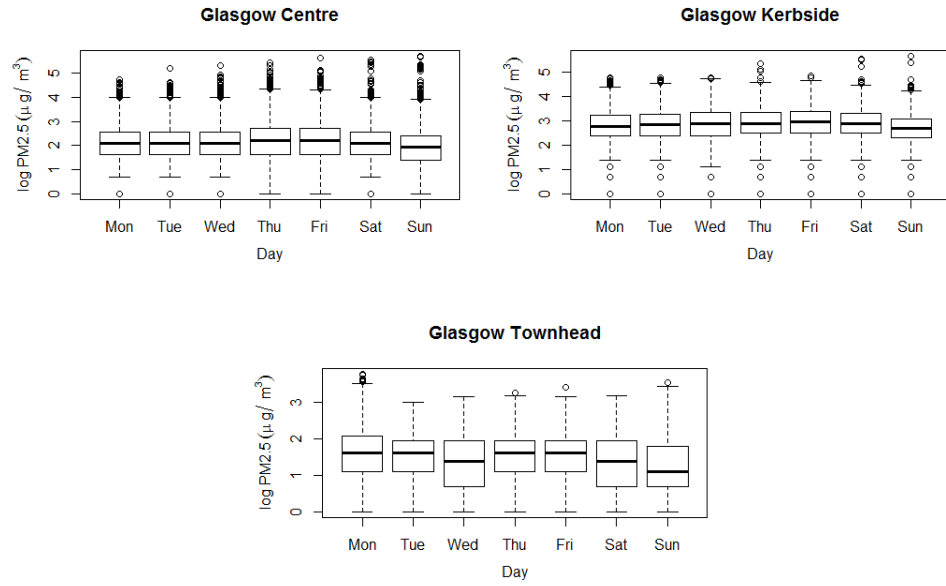


Figure A7: Boxplots of day-within-week log-transformed  $PM_{2.5}$  concentrations ( $\mu\text{g}/\text{m}^3$ ) across all monitoring stations

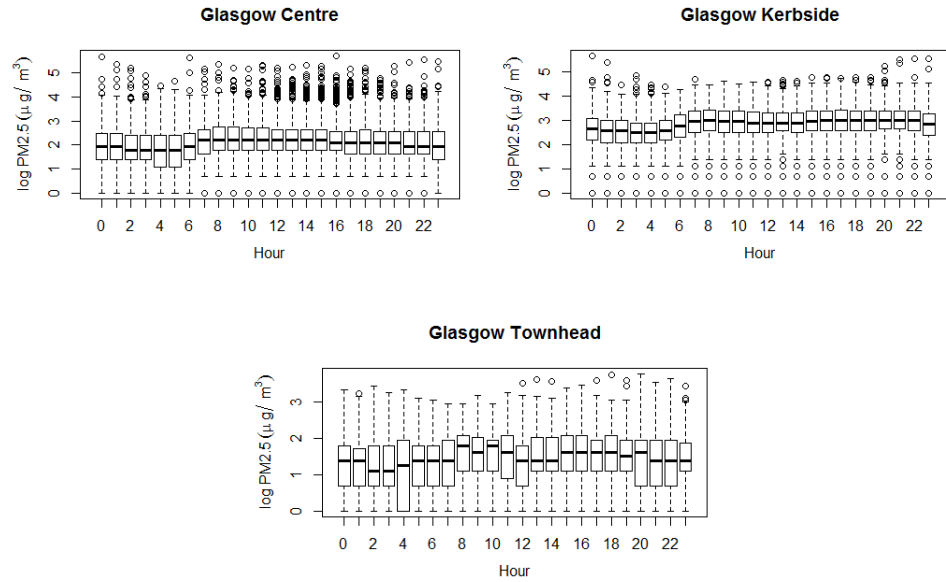


Figure A8: Boxplots of hourly log-transformed  $PM_{2.5}$  concentrations ( $\mu\text{g}/\text{m}^3$ ) across all monitoring stations



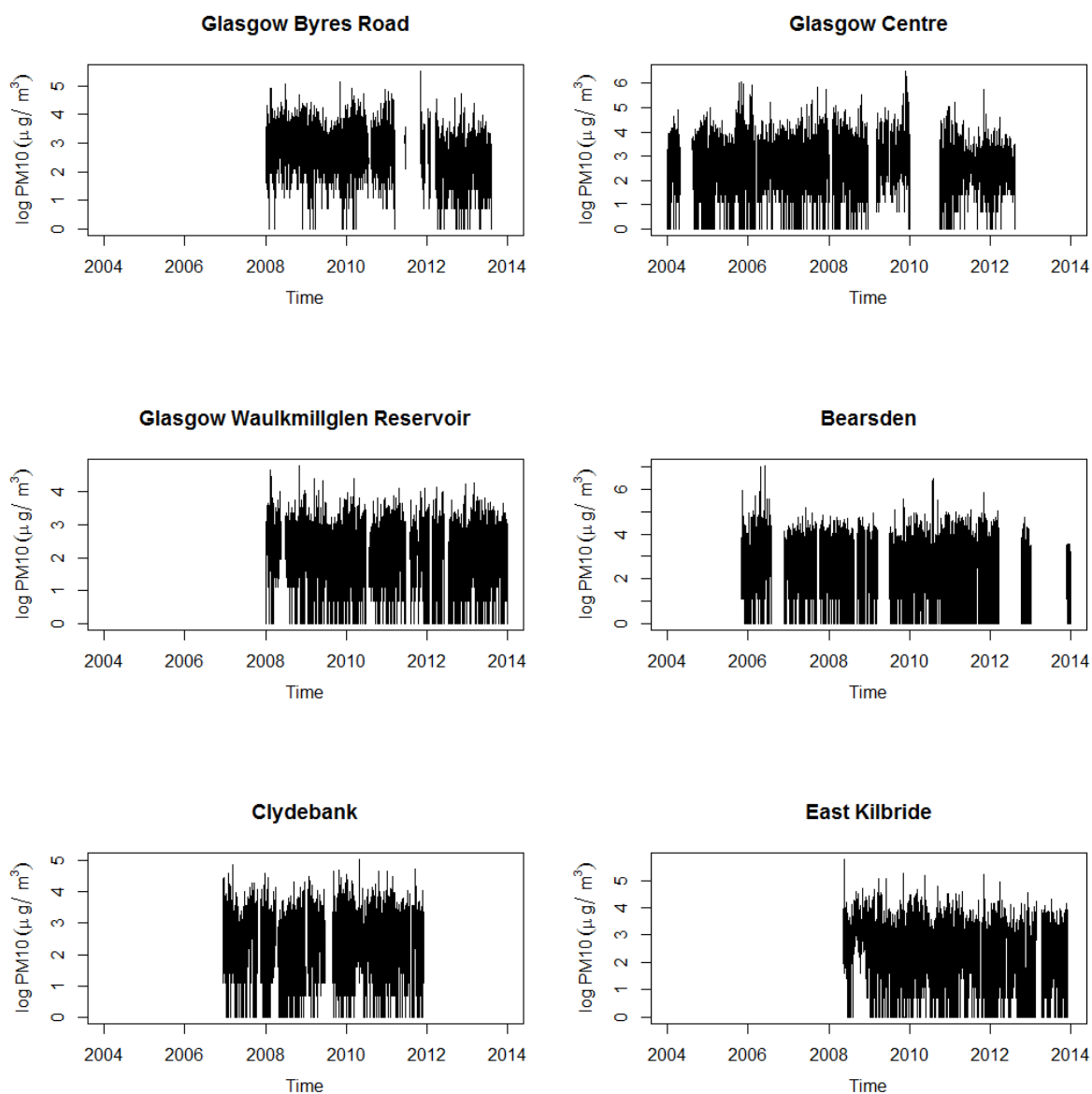


Figure A9: Time series plots of log-transformed  $PM_{10}$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

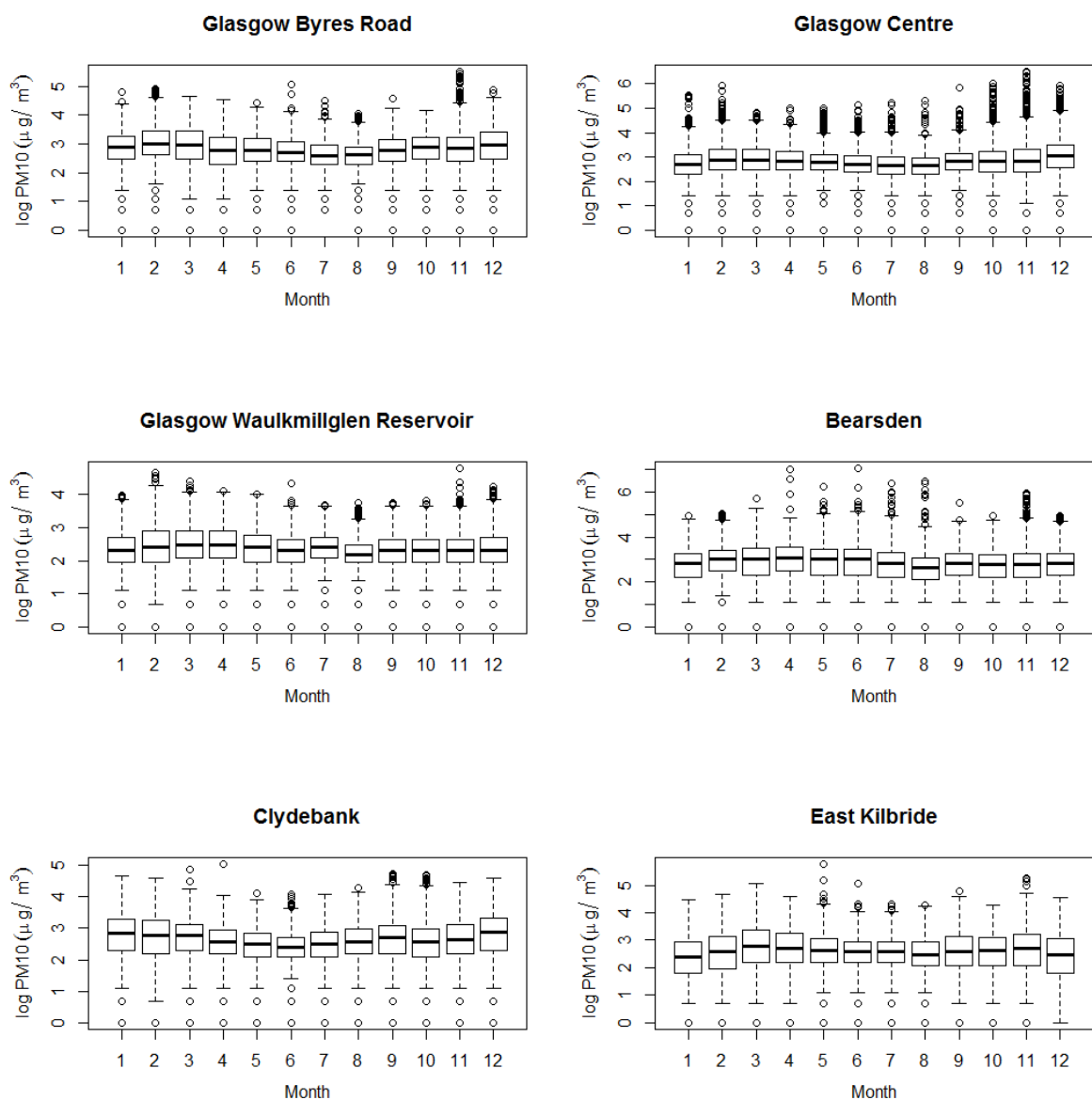


Figure A10: Boxplots of monthly log-transformed  $PM_{10}$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

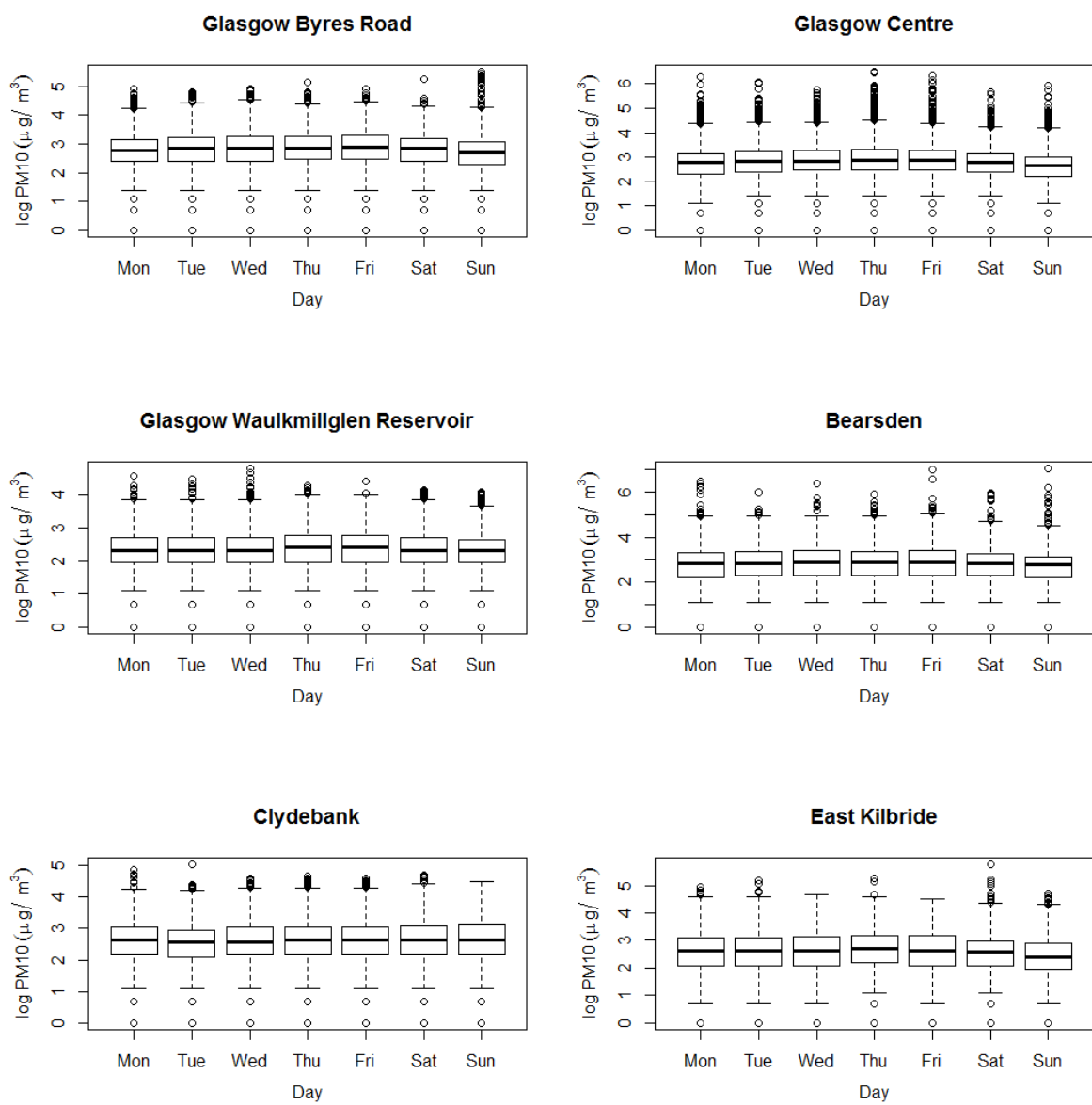


Figure A11: Boxplots of day-within-week log-transformed  $PM_{10}$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

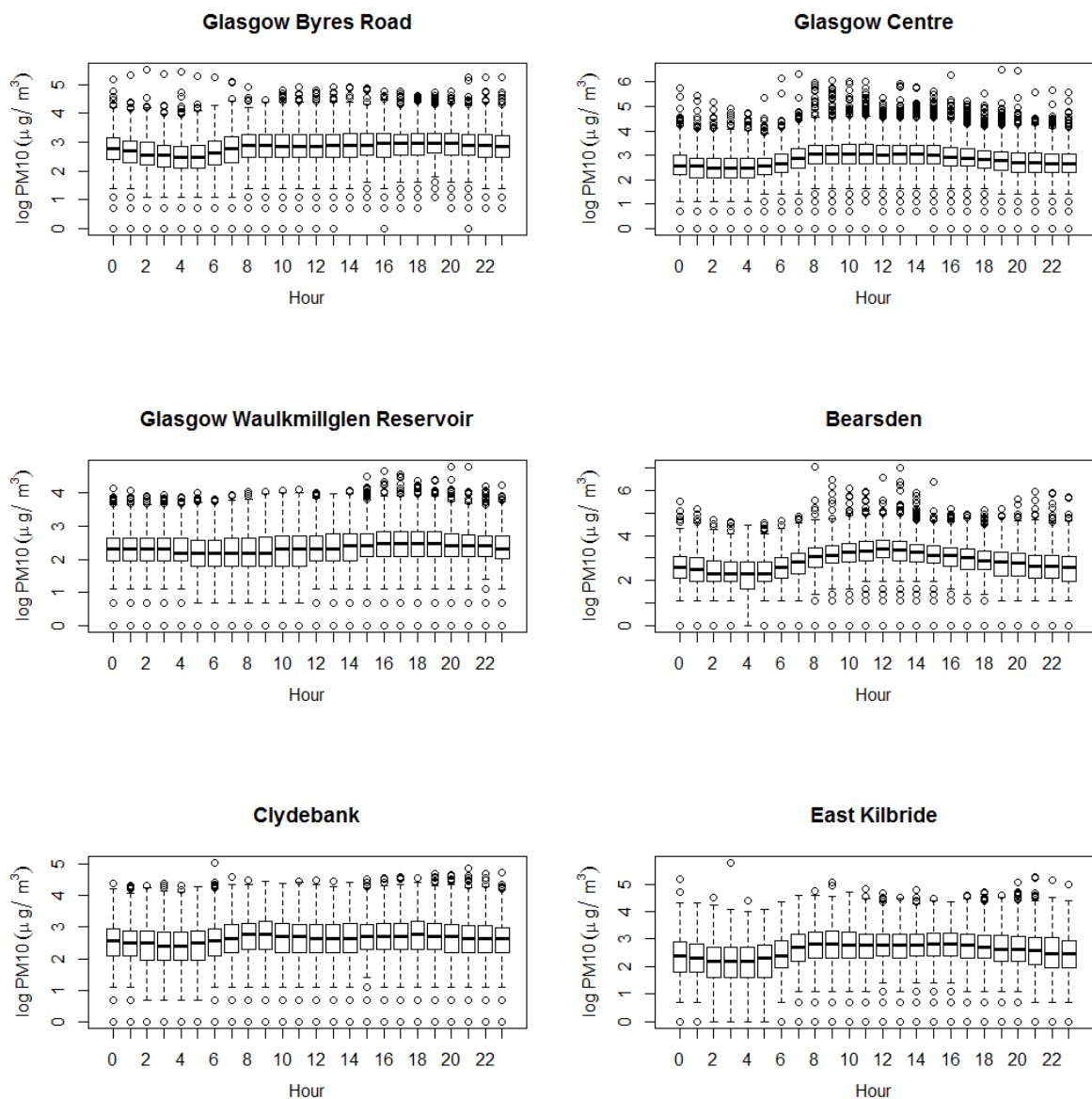


Figure A12: Boxplots of hourly log-transformed  $PM_{10}$  concentrations ( $\mu g/m^3$ ) at selected monitoring stations

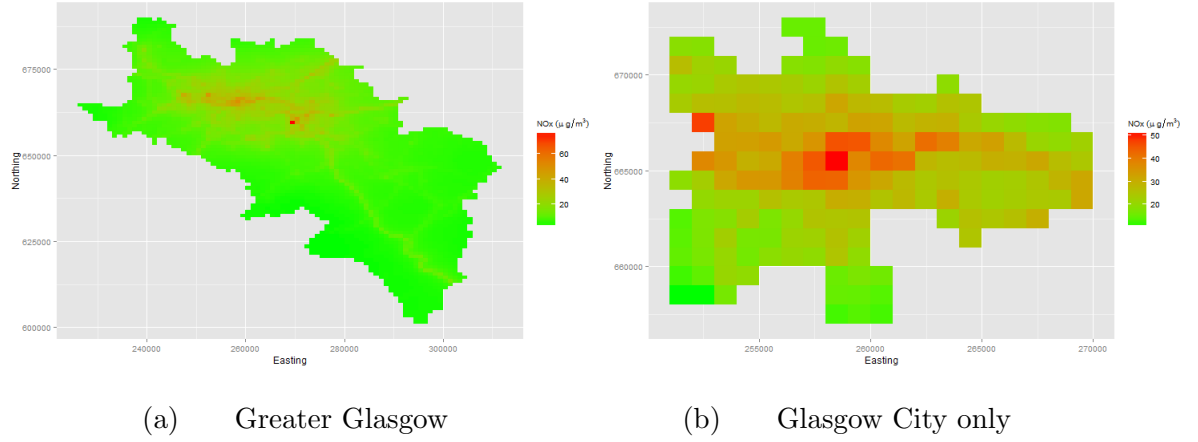


Figure A13: Maps with average 2011  $NO_x$  concentrations ( $\mu g/m^3$ ) for each 1km by 1km grid

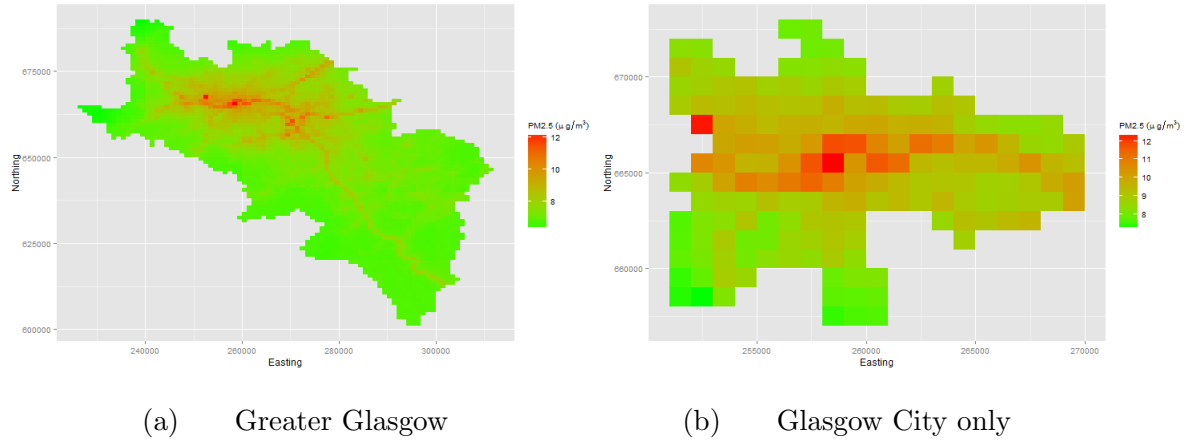


Figure A14: Maps with average 2011  $PM_{2.5}$  concentrations ( $\mu g/m^3$ ) for each 1km by 1km grid

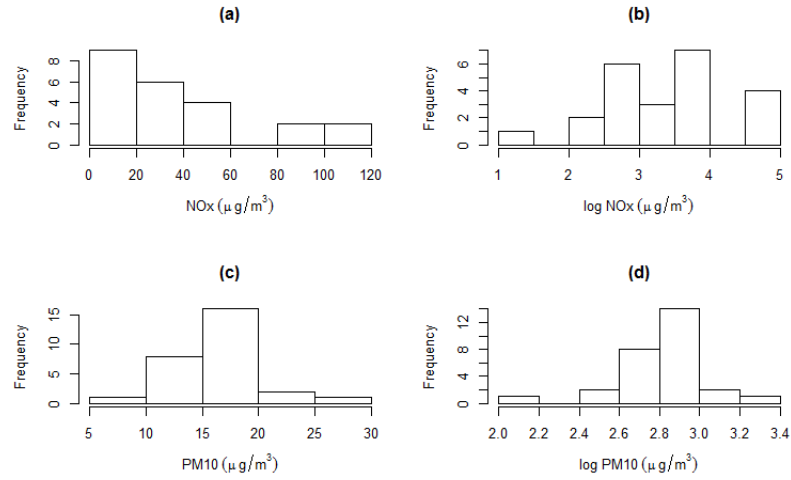


Figure A15: (a) histogram of  $NO_X$  concentration ( $\mu\text{g}/\text{m}^3$ ) values; (b) histogram of log-transformed  $NO_X$  concentration ( $\mu\text{g}/\text{m}^3$ ) values; (c) histogram of  $PM_{10}$  concentration ( $\mu\text{g}/\text{m}^3$ ) values; (d) histogram of log-transformed  $PM_{10}$  concentration ( $\mu\text{g}/\text{m}^3$ ) values (all histograms show monitoring station data only;  $PM_{2.5}$  is not included since there were very few active monitoring stations in 2011)

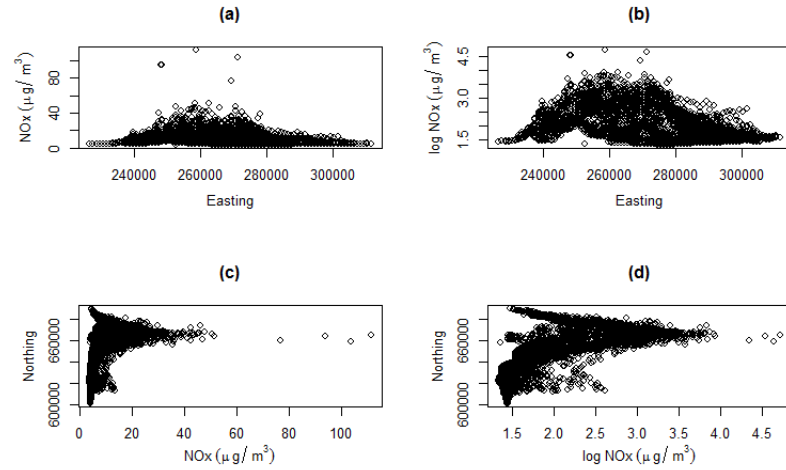


Figure A16: (a) scatterplot of easting versus  $NO_X$  concentration ( $\mu\text{g}/\text{m}^3$ ); (b) scatterplot of easting versus log-transformed  $NO_X$  concentration ( $\mu\text{g}/\text{m}^3$ ); (c) scatterplot of northing versus  $NO_X$  concentration ( $\mu\text{g}/\text{m}^3$ ); (d) scatterplot of northing versus log-transformed  $NO_X$  concentration ( $\mu\text{g}/\text{m}^3$ )

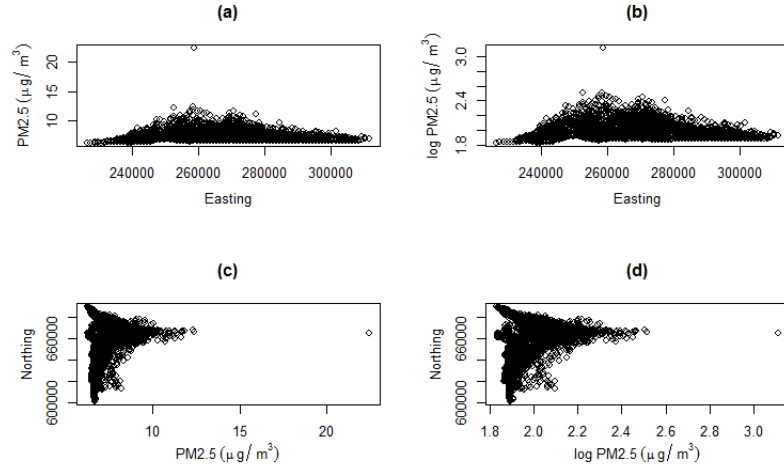


Figure A17: (a) scatterplot of easting versus  $PM_{2.5}$  concentration ( $\mu\text{g}/\text{m}^3$ ); (b) scatterplot of easting versus log-transformed  $PM_{2.5}$  concentration ( $\mu\text{g}/\text{m}^3$ ); (c) scatterplot of northing versus  $PM_{2.5}$  concentration ( $\mu\text{g}/\text{m}^3$ ); (d) scatterplot of northing versus log-transformed  $PM_{2.5}$  concentration ( $\mu\text{g}/\text{m}^3$ )

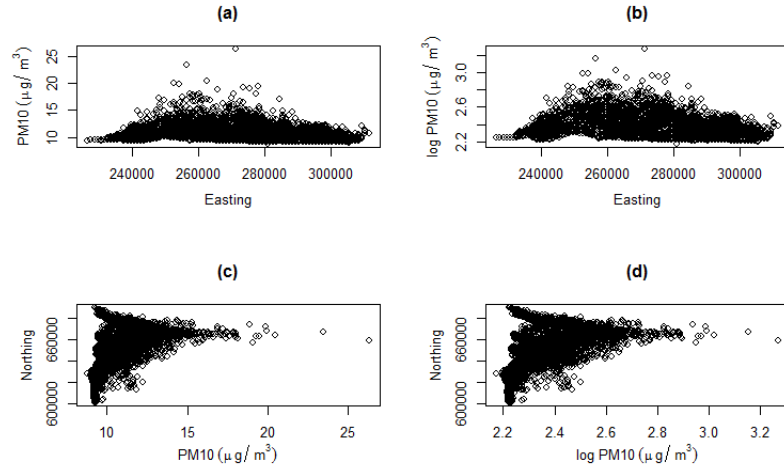


Figure A18: (a) scatterplot of easting versus  $PM_{10}$  concentration ( $\mu\text{g}/\text{m}^3$ ); (b) scatterplot of easting versus log-transformed  $PM_{10}$  concentration ( $\mu\text{g}/\text{m}^3$ ); (c) scatterplot of northing versus  $PM_{10}$  concentration ( $\mu\text{g}/\text{m}^3$ ); (d) scatterplot of northing versus log-transformed  $PM_{10}$  concentration ( $\mu\text{g}/\text{m}^3$ )

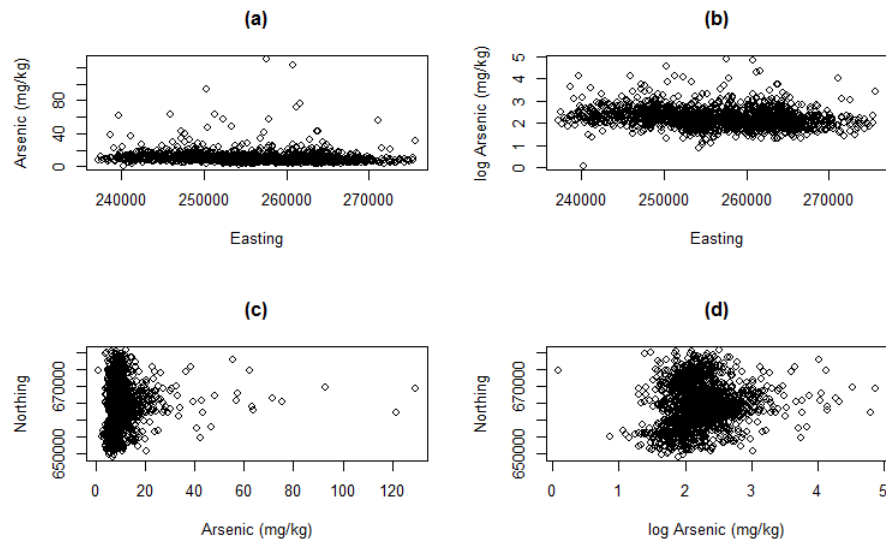


Figure A19: (a) scatterplot of easting versus arsenic concentration (mg/kg); (b) scatterplot of easting versus log-transformed arsenic concentration (mg/kg); (c) scatterplot of northing versus arsenic concentration (mg/kg); (d) scatterplot of northing versus log-transformed arsenic concentration (mg/kg)

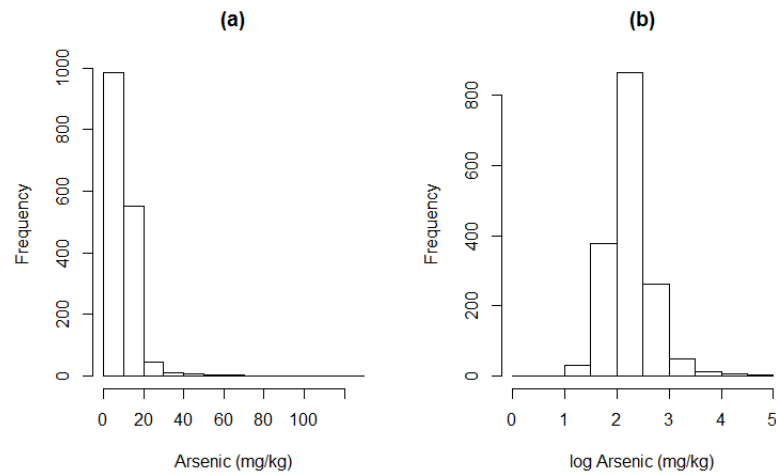


Figure A20: (a) histogram of arsenic concentration (mg/kg) values; (b) histogram of log-transformed arsenic concentration (mg/kg) values



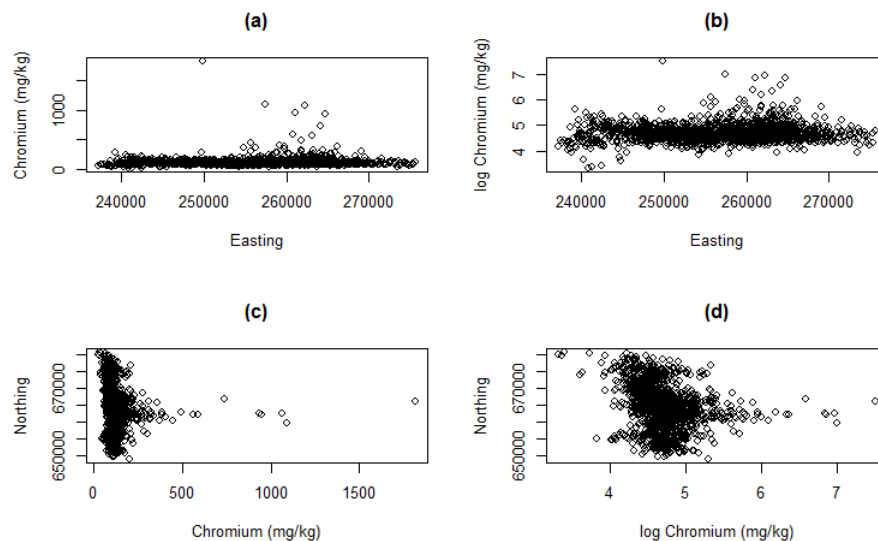


Figure A21: (a) scatterplot of easting versus chromium concentration (mg/kg); (b) scatterplot of easting versus log-transformed chromium concentration (mg/kg); (c) scatterplot of northing versus chromium concentration (mg/kg); (d) scatterplot of northing versus log-transformed chromium concentration (mg/kg)

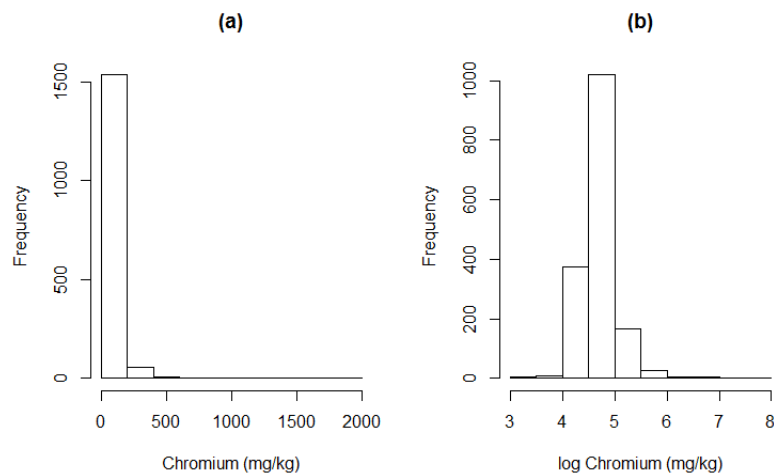


Figure A22: (a) histogram of chromium concentration values (mg/kg); (b) histogram of log-transformed chromium concentration values (mg/kg)

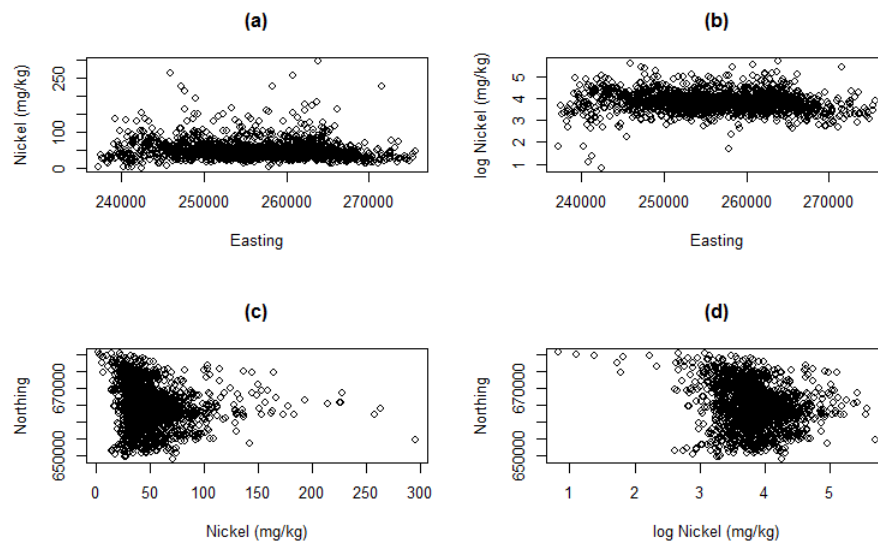


Figure A23: (a) scatterplot of easting versus nickel concentration (mg/kg); (b) scatterplot of easting versus log-transformed nickel concentration (mg/kg); (c) scatterplot of northing versus nickel concentration (mg/kg); (d) scatterplot of northing versus log-transformed nickel concentration (mg/kg)

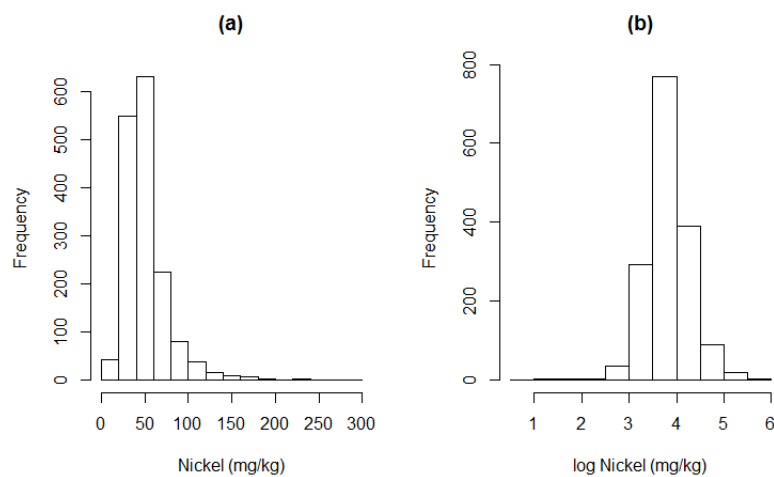


Figure A24: (a) histogram of nickel concentration (mg/kg) values; (b) histogram of log-transformed nickel concentration (mg/kg) values

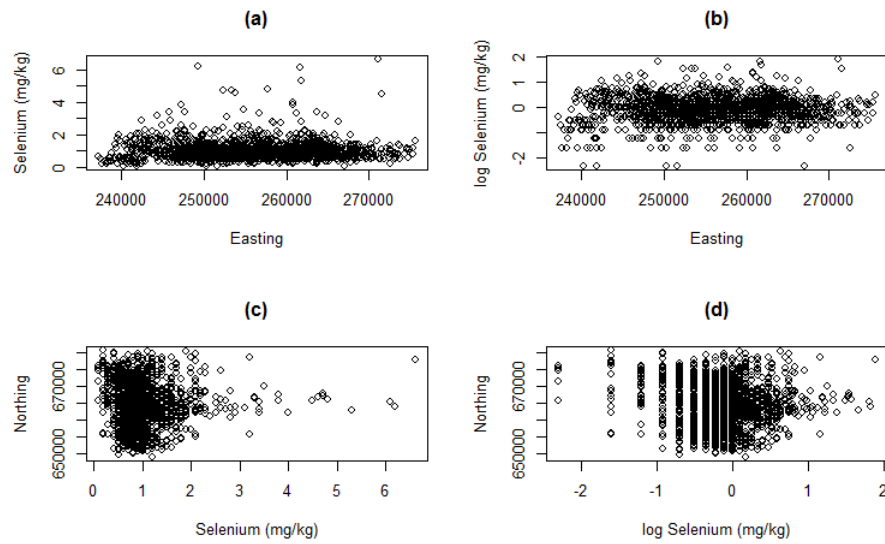


Figure A25: (a) scatterplot of easting versus selenium concentration (mg/kg); (b) scatterplot of easting versus log-transformed selenium concentration (mg/kg); (c) scatterplot of northing versus selenium concentration (mg/kg); (d) scatterplot of northing versus log-transformed selenium concentration (mg/kg)

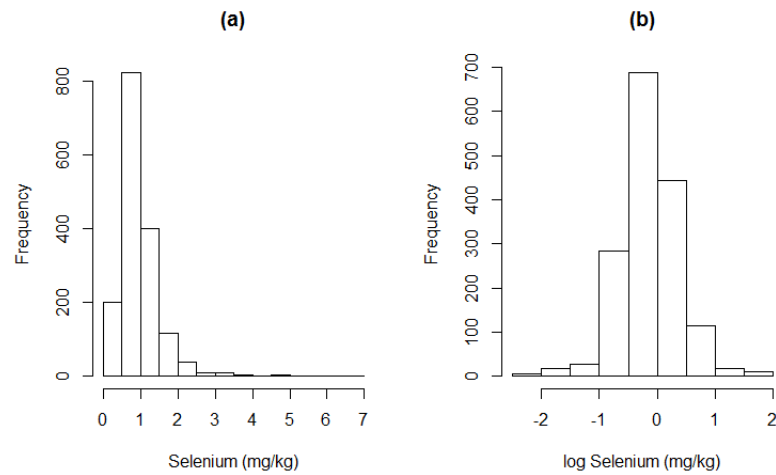


Figure A26: (a) histogram of selenium concentration (mg/kg) values; (b) histogram of log-transformed selenium concentration (mg/kg) values

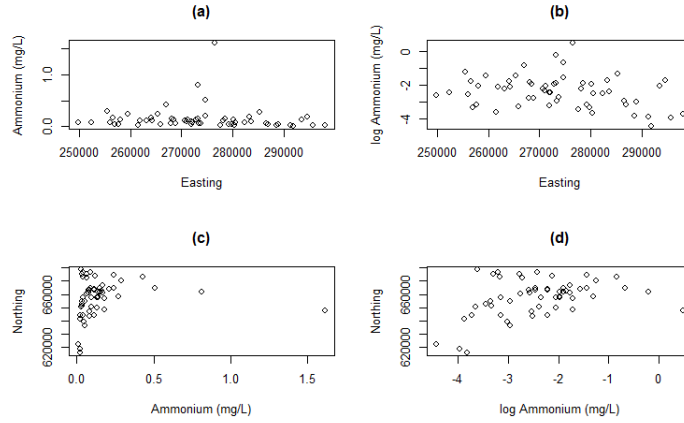


Figure A27: (a) plot of Easting versus ammonium concentration (mg/L); (b) plot of Easting versus log-transformed ammonium concentration (mg/L); (c) plot of ammonium concentration (mg/L) versus Northing; (d) plot of log-transformed ammonium concentration (mg/L) versus Northing

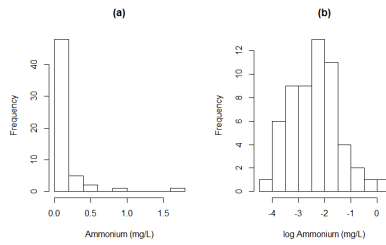


Figure A28: (a) boxplots of annual log-transformed ammonium concentration (mg/L); (b) boxplots of monthly log-transformed ammonium concentration (mg/L)

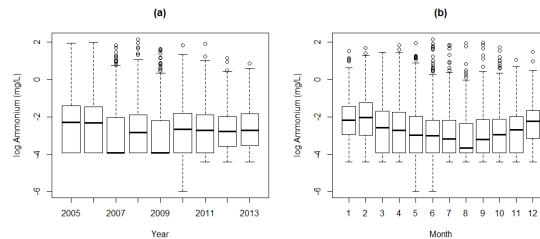


Figure A29: (a) boxplots of annual log-transformed ammonium concentration (mg/L); (b) boxplots of monthly log-transformed ammonium concentration (mg/L)

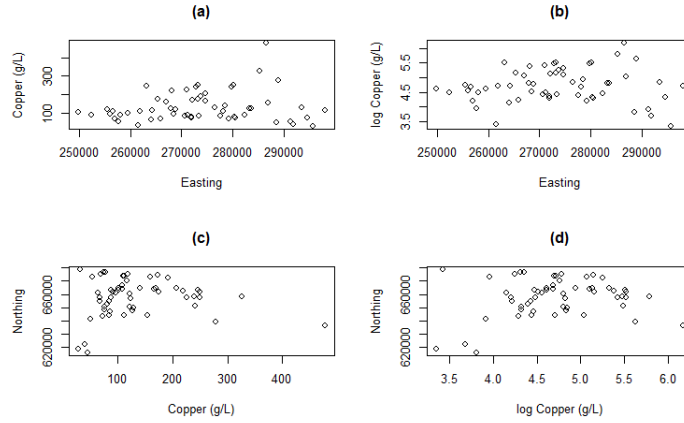


Figure A30: (a) plot of Easting versus copper concentration (g/L); (b) plot of Easting versus log-transformed copper concentration (g/L); (c) plot of copper concentration (g/L) versus Northing; (d) plot of log-transformed copper concentration (g/L) versus Northing

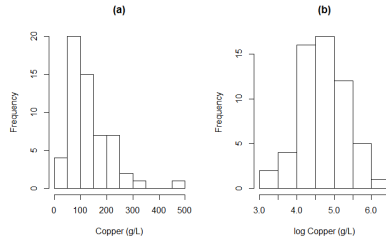


Figure A31: (a) boxplots of annual log-transformed copper concentration (g/L); (b) boxplots of monthly log-transformed copper concentration (g/L)

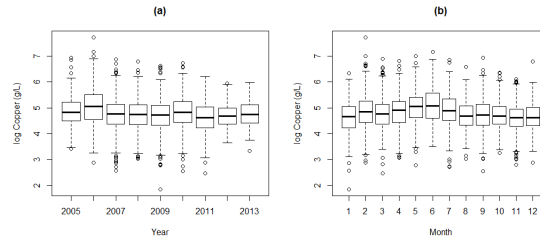


Figure A32: (a) boxplots of annual log-transformed copper concentration (g/L); (b) boxplots of monthly log-transformed copper concentration (g/L)

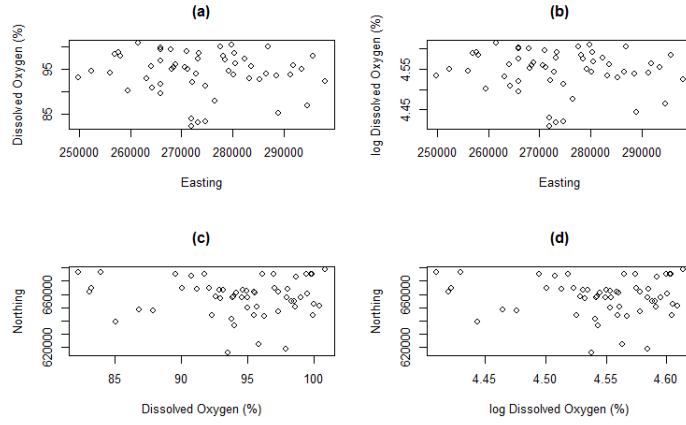


Figure A33: (a) plot of Easting versus dissolved oxygen concentration (%); (b) plot of Easting versus log-transformed dissolved oxygen concentration (%); (c) plot of dissolved oxygen concentration (%) versus Northing; (d) plot of log-transformed dissolved oxygen concentration (%) versus Northing

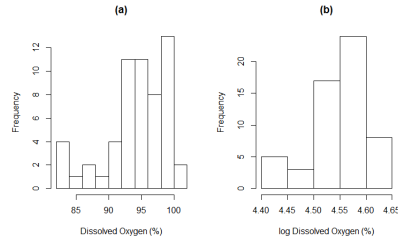


Figure A34: (a) boxplots of annual log-transformed dissolved oxygen concentration (%); (b) boxplots of monthly log-transformed dissolved oxygen concentration (%)

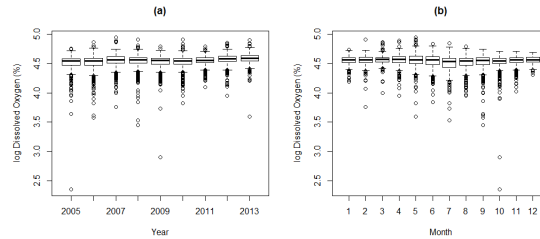


Figure A35: (a) boxplots of annual log-transformed dissolved oxygen concentration (%); (b) boxplots of monthly log-transformed dissolved oxygen concentration (%)

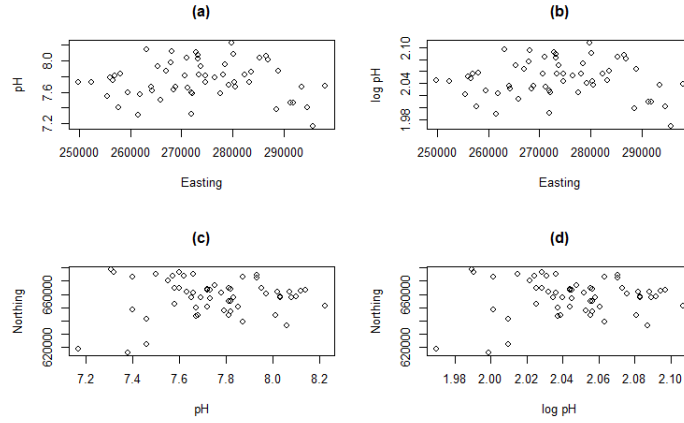


Figure A36: (a) plot of Easting versus pH concentration; (b) plot of Easting versus log-transformed pH concentration; (c) plot of pH concentration versus Northing; (d) plot of log-transformed pH concentration versus Northing

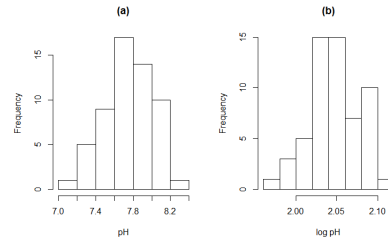


Figure A37: (a) boxplots of annual log-transformed pH concentration; (b) boxplots of monthly log-transformed pH concentration

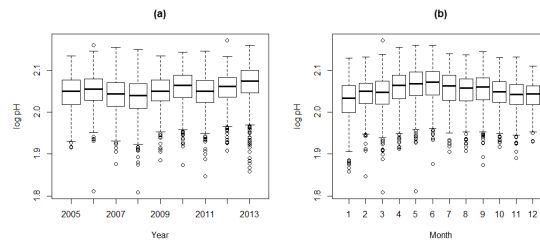


Figure A38: (a) boxplots of annual log-transformed pH concentration; (b) boxplots of monthly log-transformed pH concentration

## B. Additional Material Related to Chapter 3

Table B1: Results of flexible regression model for  $NO_X$  ( $k = 575$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	2.07	0.0034	610.7	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	28.72	28.99	901.8	< 0.05

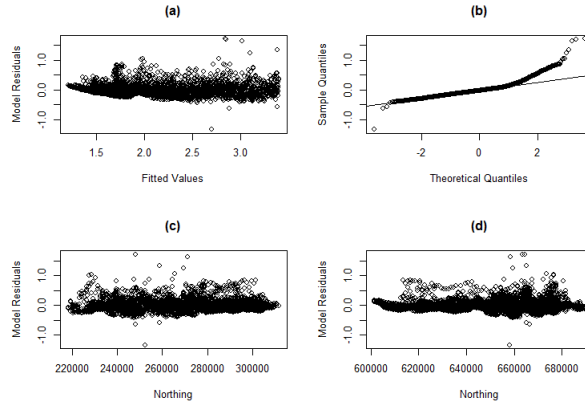


Figure B1: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

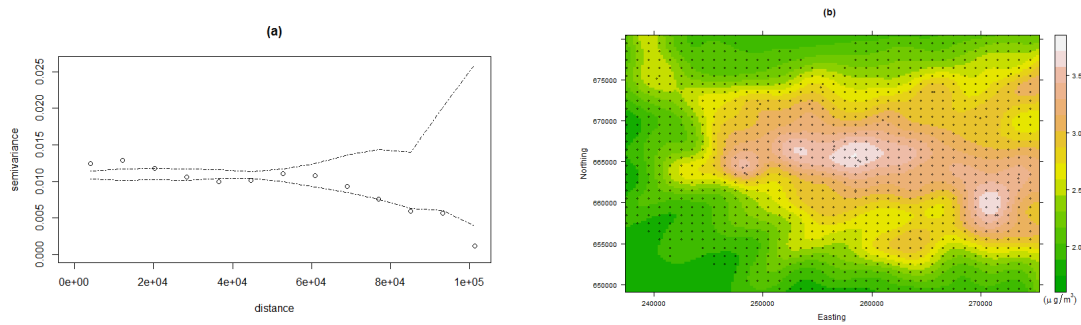


Figure B2: (a) variogram of residuals; (b) model prediction surface with exponential correlation structure (log-transformed  $NO_X$  concentration –  $\mu g/m^3$ )



Table B2: Results of flexible regression model for  $PM_{2.5}$  ( $k = 525$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	1.98	0.00084	2351	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	28.7	28.99	477.2	< 0.05

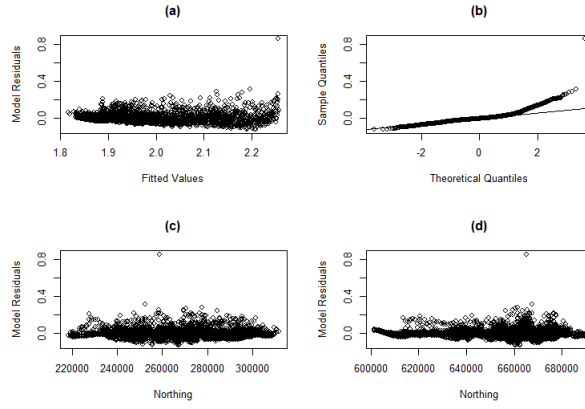


Figure B3: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

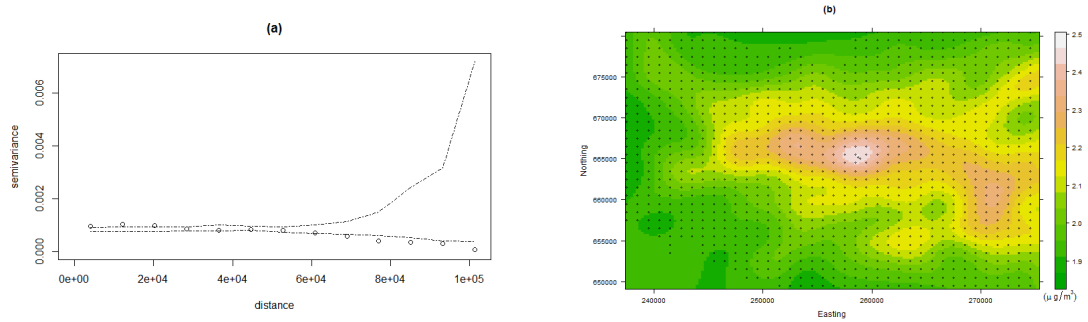


Figure B4: (a) variogram of residuals; (b) model prediction surface with exponential correlation structure (log-transformed  $PM_{2.5}$  concentration –  $\mu\text{g}/\text{m}^3$ )

Table B3: Results of flexible regression model for  $PM_{10}$  ( $k = 575$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	2.38	0.0012	2069	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	28.62	28.99	364.4	< 0.05

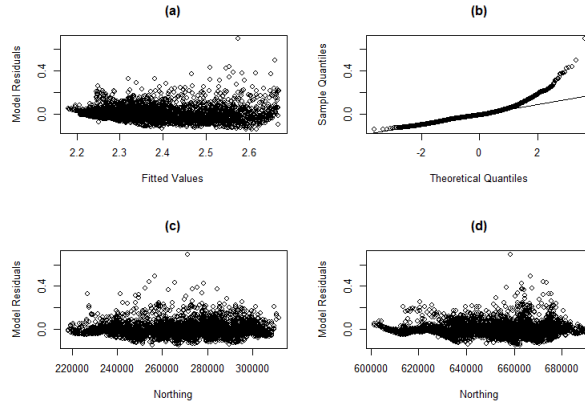


Figure B5: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

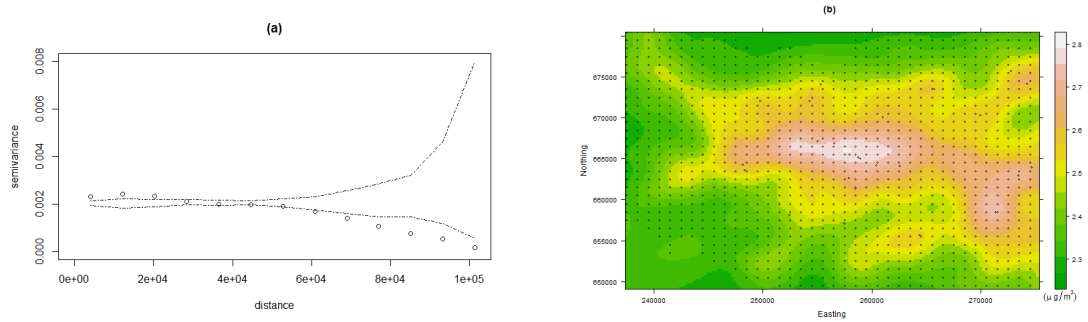


Figure B6: (a) variogram of residuals; (b) model prediction surface with exponential correlation structure (log-transformed  $PM_{10}$  concentration –  $\mu\text{g}/\text{m}^3$ )

Table B4: Results of flexible regression model for arsenic ( $k = 70$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	2.247	0.01	224.9	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	25.62	28.32	11.14	< 0.05

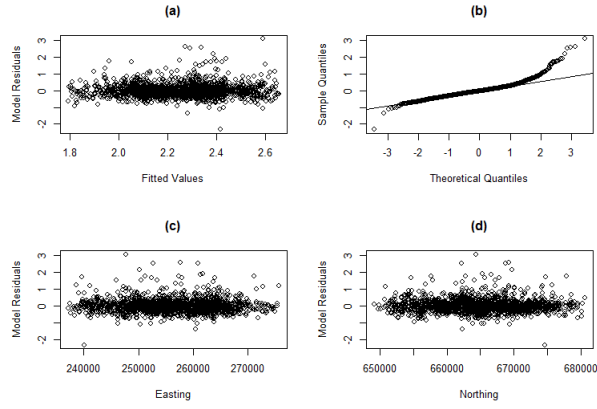


Figure B7: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

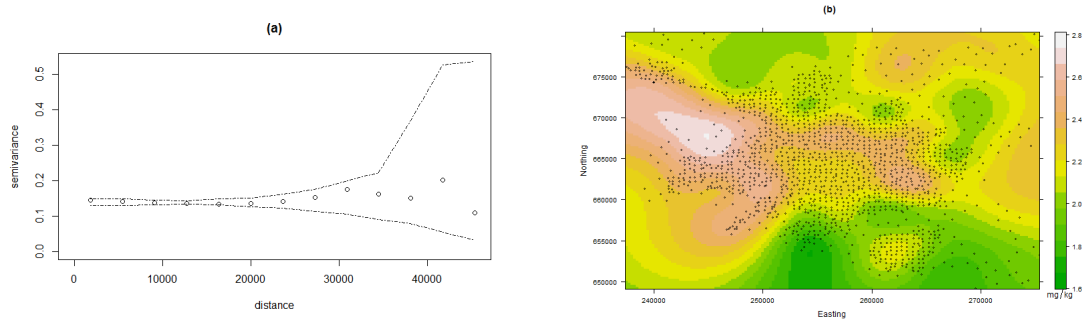


Figure B8: (a) variogram of residuals; (b) model prediction surface (log-transformed arsenic concentration – mg/kg)

Table B5: Results of flexible regression model for chromium ( $k = 110$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	4.705	0.008	616.5	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	25.09	28.11	19.23	< 0.05

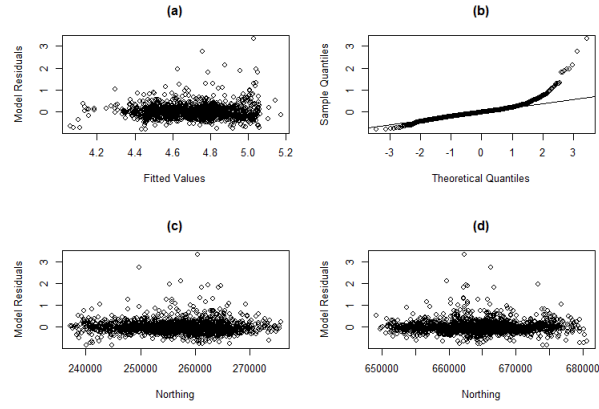


Figure B9: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

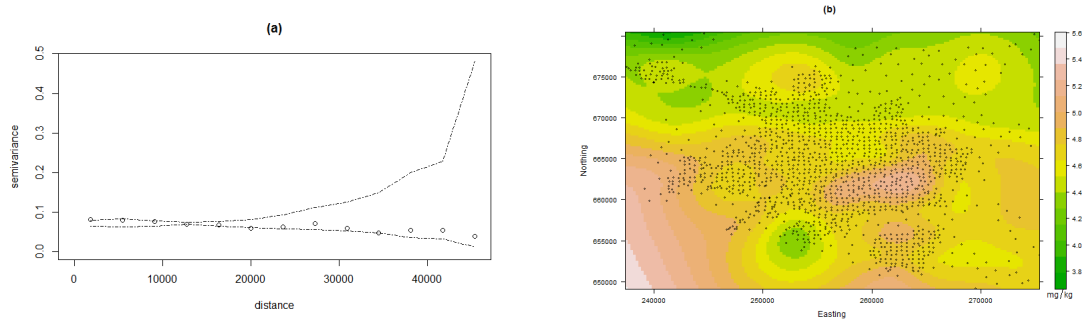


Figure B10: (a) variogram of residuals; (b) model prediction surface (log-transformed chromium concentration – mg/kg)

Table B6: Results of flexible regression model for nickel ( $k = 130$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	3.836	0.01	370.1	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	25.4	28.24	17.95	< 0.05

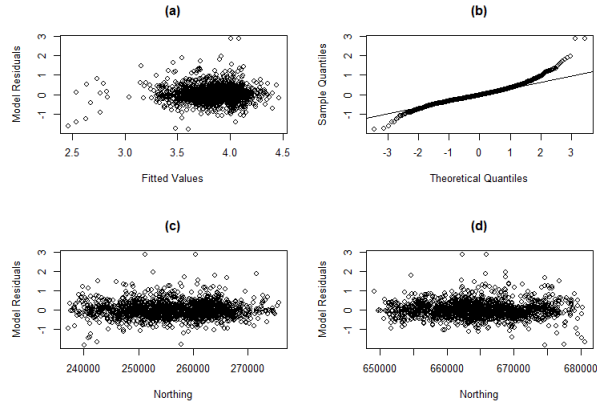


Figure B11: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

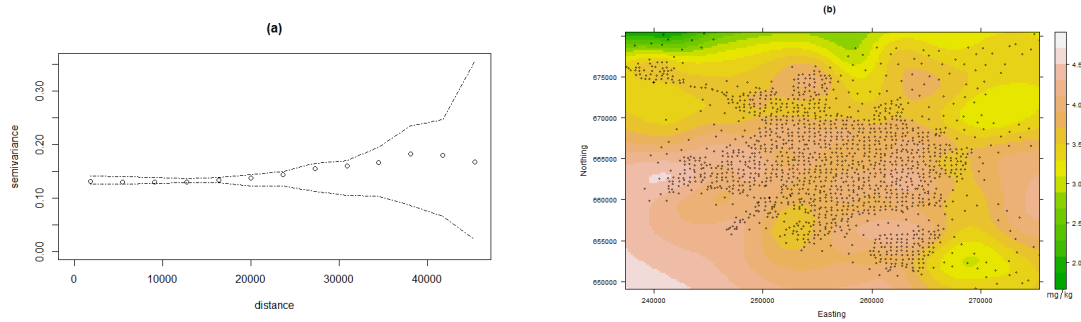


Figure B12: (a) variogram of residuals; (b) model prediction surface (log-transformed nickel concentration – mg/kg)

Table B7: Results of flexible regression model for selenium ( $k = 160$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	-0.095	0.012	-8.17	< 0.05
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	20.8	25.44	9.33	< 0.05

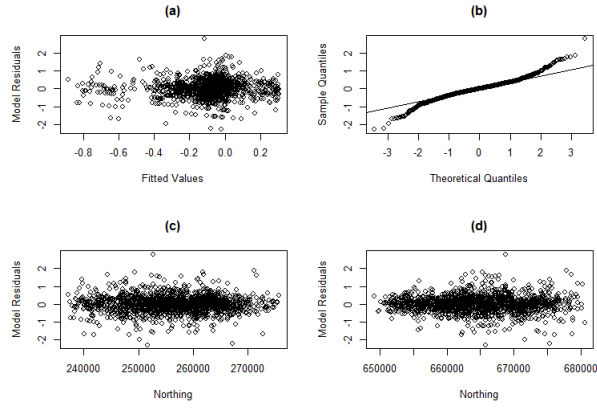


Figure B13: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

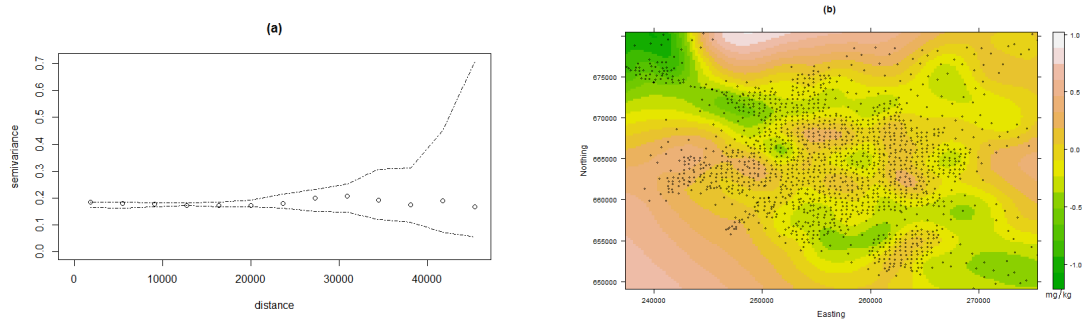


Figure B14: (a) variogram of residuals; (b) model prediction surface (log-transformed selenium concentration – mg/kg)

Table B8: Results of flexible regression model for ammonium ( $k = 15$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	-2.06	0.15	-13.66	< 0.05
Catchment (Kelvin)	-1.08	0.36	-3.04	< 0.05
Catchment (Cart)	-1.00	0.53	-1.88	0.06
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	2.00	2.00	10.59	< 0.05

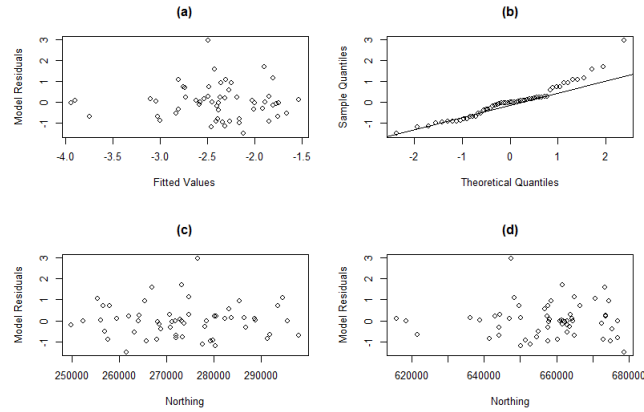


Figure B15: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

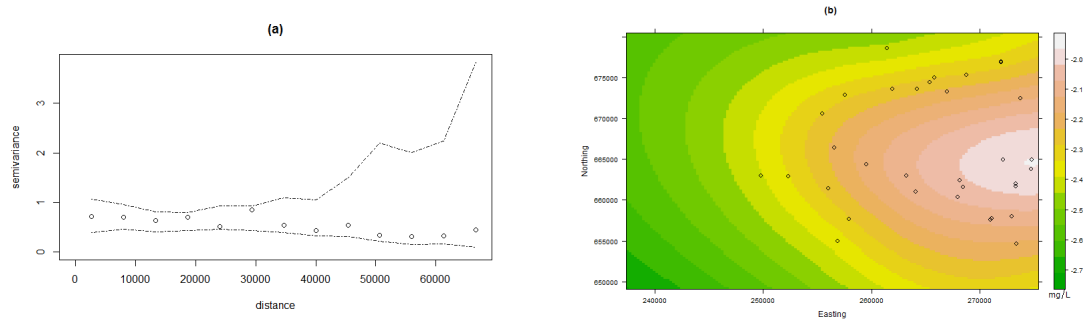


Figure B16: (a) variogram of residuals; (b) model prediction surface (log-transformed ammonium concentration – mg/L)

Table B9: Results of flexible regression model for copper ( $k = 15$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	4.72	0.13	35.16	< 0.05
Catchment (Kelvin)	0.06	0.43	0.13	0.90
Catchment (Cart)	-0.06	0.48	-0.12	0.90
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	8.24	10.58	2.54	< 0.05

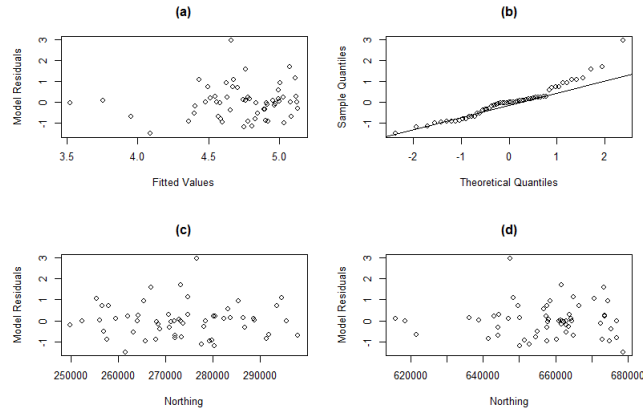


Figure B17: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

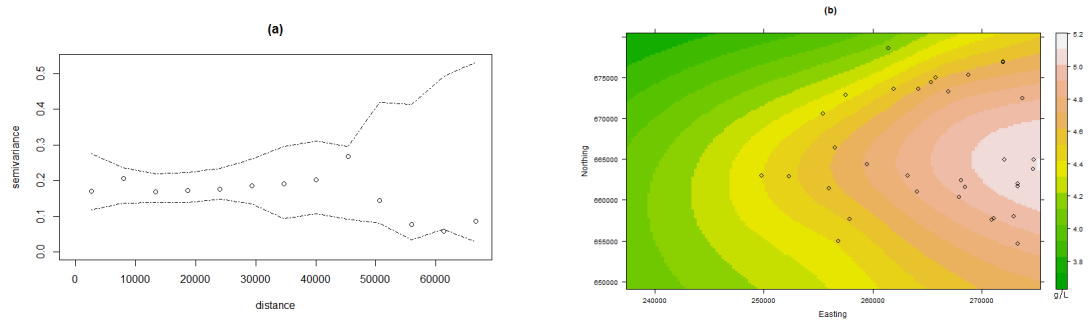


Figure B18: (a) variogram of residuals; (b) model prediction surface (log-transformed copper concentration – g/L)



Table B10: Results of flexible regression model for dissolved oxygen ( $k = 15$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	4.54	0.01	444.39	< 0.05
Catchment (Kelvin)	0.03	0.04	0.65	0.52
Catchment (Cart)	-0.01	0.04	-0.30	0.77
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	4.57	6.12	0.84	0.55

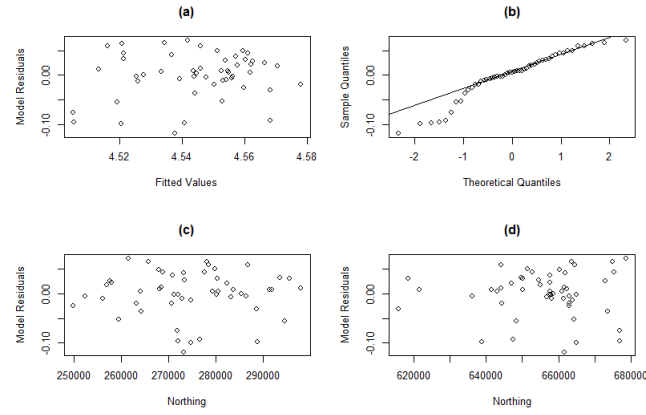


Figure B19: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

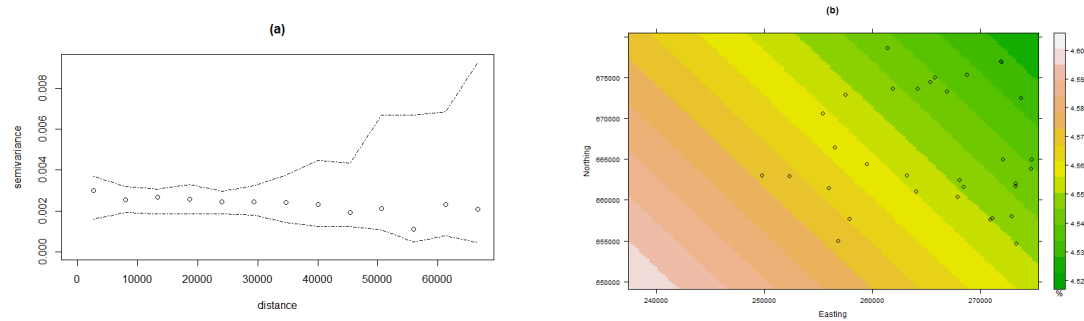


Figure B20: (a) variogram of residuals; (b) model prediction surface (log-transformed dissolved oxygen concentration – % saturation)

Table B11: Results of flexible regression model for pH ( $k = 15$ )

Parametric Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	2.05	0.006	321.17	< 0.05
Catchment (Kelvin)	0.002	0.02	0.1	0.92
Catchment (Cart)	0.01	0.02	0.34	0.73
Smooth Coefficient	Effective Degrees of Freedom	Reference Degrees of Freedom	F-value	p-value
Easting : Northing	6.96	9.20	3.14	< 0.05

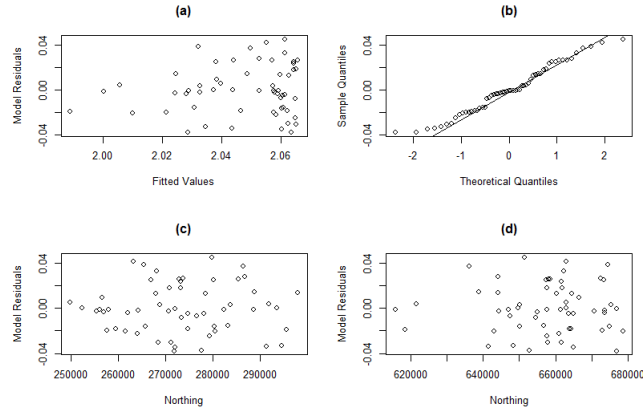


Figure B21: (a) scatterplot of model fitted values versus model residuals; (b) normal Q-Q plot of model residuals; (c) scatterplot of easting versus model residuals; (d) scatterplot of northing versus model residuals

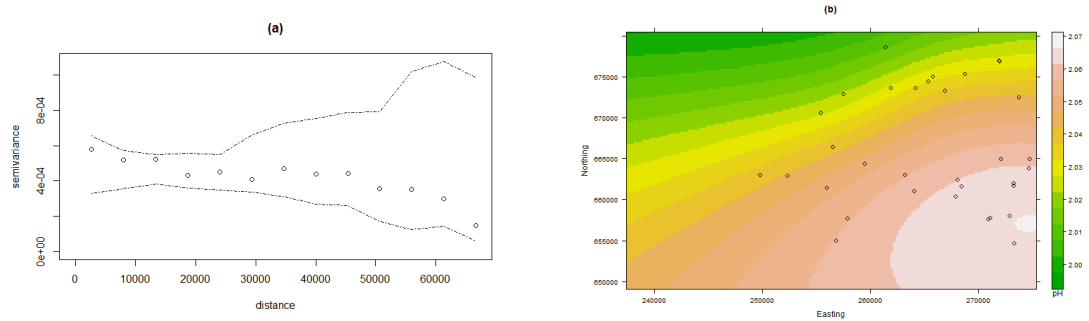


Figure B22: (a) variogram of residuals; (b) model prediction surface (log-transformed pH concentration)

# Bibliography

Ahuja, S. (2013). *Monitoring Water Quality: Pollution Assessment, Analysis and Remediation*. Boston: Elsevier.

Air Quality in Scotland (2014). <http://www.scottishairquality.co.uk>.

Akaike, H. (1974). “A new look at the statistical model identification.” *IEEE Transactions on Automatic Control*. 19 (6), 716 – 723.

Benedini, M. and Tsakiris, G. (2013). *Water Quality Modelling for Rivers and Streams*. Dordrecht: Springer.

Berkowitz, B., Dror, I. and Yaron, B. (2014). *Contaminant Geochemistry: Interactions and Transport in the Subsurface Environment*. 2<sup>nd</sup> ed. Heidelberg: Springer.

Breiman, L. (1996). “Bagging Predictors.” *Machine Learning*. 24, 123 – 140.

Breiman, L. (2001). “Random Forests.” *Machine Learning*. 45, 5 – 32.

Breiman, L. and Cutler, A. (no date). [http://www.stat.berkeley.edu/~breiman/RandomForests/cc\\_home.htm](http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm).

Bruno, F. and Cocchi, D. (2002). “A unified strategy for building simple air quality indices.” *Environmetrics*. 13, 243 – 261.

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodal Inference: A Practical Information-Theoretical Approach*. 2<sup>nd</sup> ed. Springer.

Canadian Council of Ministers of the Environment (2001). *Canadian water quality guidelines for the protection of aquatic life*. CCME Water Quality Index 1.0, Technical Report. In: Canadian environmental quality guidelines, 1999, Canadian Council of Ministers of the Environment, Winnipeg.

Canadian Council of Ministers of the Environment (2007). CCME Soil Quality Index 1.0: Technical Report. In: Canadian environmental quality guidelines, 1999, Canadian Council of Ministers of the Environment, Winnipeg.

Cattell, R.B. (1966). “The scree test for the number of factors.” *Multivariate Behavioural Research*. I, 245 – 276.

Chiras, D.P. (2006). *Environmental Science*. 7<sup>th</sup> ed. London: Jones and Bartlett Learning.

Cohen, J. (1960). “A coefficient of agreement for nominal scales.” *Educational and Psychological Measurement*. 20 (1), 37 – 46.

Cohen, J. (1968). “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.” *Psychological Bulletin*. 70, 213 – 220.

Cutler, A. (2010). Random Forests for Regression and Classification.  
<http://www.math.usu.edu/adele/randomforests/ovornnaz.pdf>.

Diggle, P.J., Menezes, R. and Su, T. (2010). “Geostatistical inference under preferential sampling.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 59 (2), 191 – 232.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-Based Geostatistics*. New York, NY: Springer.

Dobbie, J.M. and Clifford, D. (2015). “Quantifying uncertainty in environmental indices: an application to an estuarine health index.” *Marine and Freshwater Research*. 60, 95 – 105.

Dobbie, K.E., Bruneau, P.M.C. and Towers, W. (eds). (2011). *The State of Scotland’s Soil*. Natural Scotland. [www.sepa.org.uk/land/land\\_publications.aspx](http://www.sepa.org.uk/land/land_publications.aspx).

Doncaster Metropolitan Borough Council (2014).  
[http://www.doncaster.gov.uk/airq/testing\\_air\\_quality/diffusion\\_tubes.asp](http://www.doncaster.gov.uk/airq/testing_air_quality/diffusion_tubes.asp).

European Commission (2014).  
[http://www.ec.europa.eu/environment/soil/index\\_en.htm](http://www.ec.europa.eu/environment/soil/index_en.htm).

European Environment Agency (2013). *Air Quality in Europe – 2013 Report*.  
<http://www.eea.europa.eu/publications/air-quality-in-europe-2013>.

European Union (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. <http://eur-lex.europa.eu/legal-content/EN/NOT/?uri=CELEX:32000L0060>.

Eurostat (2014). *Towards a harmonised methodology for statistical indicators – Part 1: Indicator typologies and terminologies*. <http://ec.europa.eu/eurostat/en/web/products-manuals-and-guidelines/-/KS-GQ-14-011>.

Fahrmeir, L., Lang, S., Kneib, T. and Marx, B. (2013). *Regression: Models, Methods and Applications*. Berlin-Heidelberg: Springer.

Faraway, J.J. (2006). *Extending Linear Models with R: Generalised Linear, Mixed Effects and Nonparametric Regression Models*. London: Chapman & Hall.

Fleiss, J.L., Cohen, J. and Everitt, B.S. (1969). “Large sample standard errors of kappa and weighted kappa”. *Psychological Bulletin*. 72, 323 – 327.

Fordyce, F.M., Nice, S.E., Lister, T.R., Ó Dochartaigh, B.É., Cooper, R., Allen, M., Ingham, M., Gowing, C., Vickers, B.P. and Scheib, A. (2012). *Urban Soil Geochemistry of Glasgow*. Open Report, OR/08/002. British Geological Survey, Edinburgh.

Garrigues, E., Corson, M.S., Angers, D.A., van der Werf, H.M.G. and Walter, C. (2012). “Soil quality in the Life Cycle Assessment: towards development of an indicator.” *Ecological Indicators*. 18, 434 – 442.

Green, P.J. and Silverman, B.W. 1994. *Nonparametric Regression and Generalised Linear Models: a Roughness Penalty Approach*. London: Chapman & Hall.

Guerreiro, C.B.B., Foltescu, V. and de Leeuw, F. (2014). “Air quality status and trends in Europe.” *Atmospheric Environment*. 98, 376 – 384.

Hall, P. and Miller, H. (2010). “Modelling the variability of rankings.” *The Annals of Statistics*. 38 (5), 2652 – 2677.

Hardy, J.T. (2003). *Climate Change: Causes, Effects and Solutions*. Chichester: Wiley.

Hastie, T.J. (2013). Generalized additive models. R package version 1.09.1.  
<http://CRAN.R-project.org/package=gam>.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Health Protection Agency (2003). *Copper – General Information*.

[https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/318345/hpa.Copper.General.Information.v1.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/318345/hpa.Copper.General.Information.v1.pdf).

Helton, J.C. and Davis, F.J. (2000). *Sampling-Based Methods*. In: Saltelli, A., Chan, K. and Scott, E.M. *Sensitivity Analysis*. Chichester: Wiley.

Hsu, A., Reuben, A., Shindell, D., de Sherbinin, A. and Levy, M. (2013). “Toward the next generation of air quality monitoring indicators.” *Atmospheric Environment*. 80, 561 – 570.

Hsu, A., Emerson, J., Levy, M., de Sherbinin, A., Johnson, L., Malik, O., Schwartz, J. and Jaiteh, M. (2014). *The 2014 Environmental Performance Index*. New Haven, CT: Yale Centre for Environmental Law and Policy. [www.epi.yale.edu](http://www.epi.yale.edu).

Hurvich, C.M. and Tsai, C.L. (1989). “Regression and time series model selection in small samples.” *Biometrika*. 76, 297 – 307.

Jørgensen, S.E., Costanza, R. and Xu, F.L. (2005). *Handbook of Ecological Indicators for Assessment of Ecosystem Health*. Boca Raton, FL: Taylor & Francis.

Kaiser, H.F. (1960). “The application of electronic computers to factor analysis.” *Educational and Psychological Measurement*. 20, 141 – 151.

Karlen, D.L., Mausbach, M.J., Doran, J.W., Cline, R.G., Harris, R.F. and Schuman, G.E. (1997). “Soil quality: a concept, definition and framework for evaluation.” *Soil Science Society of America Journal*. 61, 4 – 10.

- Kim, J.O. and Mueller, C.W. (1978). *Factor analysis: statistical methods and practical issues*. Beverly Hills, CA: Sage.
- Leckie, G. and Goldstein, H. (2009). “The limitations of using school league tables to inform school choice.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 172 (4), 835 – 851.
- Lee, D., Ferguson, C. and Scott, M. (2011). “Constructing representative air quality indicators with measures of uncertainty.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 174 (1), 109 – 126.
- Lewis, R. and Evans, W. (2006). *Chemistry*. 3<sup>rd</sup> ed. Basingstoke: Palgrave Macmillan.
- Manolidias, O.G. (2002). “Development of ecological indicators – a methodological framework using compromise programming.” *Ecological Indicators*. 2, 169 – 176.
- Matérn, B. (1960). *Spatial Variation*: Technical Report. In: Statens Skogsforsningsinstitut, Stockholm.
- Messer, L.C., Jagai, J.S., Rappazzo, K.M. and Lobdell, D.T. (2014). “Construction of an environmental quality index for public health research.” *Environmental Health*. 13 – 39.
- Mohebbi, R.M., Saeedi, R., Montazeri, A., Vaghefi, K.A., Labbafi, S., Oktaie, S., Abtahi, M. and Mohagheghian, A. (2013). “Assessment of water quality in ground-water resources of Iran using a modified drinking water quality index (DWQI).” *Ecological Indicators*. 30, 28 – 34.



- Nardo, M., Saisana, M., Saltelli, A. and Tarantola, S. (2005). *Tools for Composite Indicator Building*. Joint Research Centre: European Commission.
- Nesaratnam, S.T. (2014). *Water Pollution Control*. Chichester: Wiley.
- Newman, M.C. (1993). “Regression analysis of log-transformed data: statistical bias and its correction.” *Environmental Toxicology and Chemistry*. 12, 1129 – 1133.
- O’Donnell, D., Rushworth, A., Bowman, A.W., Scott, E.M. and Hallard, M. (2014). “Flexible regression models over river networks.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. 63 (1), 47 – 63.
- OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. © OECD.
- Ott, W.R. (1978). *Environmental Indices: Theory and Practice*. Ann Arbor.
- Ott, W.R. and Hunt, W.F. (1976). “A quantitative evaluation of the pollutant standards index.” *Journal of the Air Pollution Control Association*. 26 (11), 1050 – 1054.
- Patterson, H.D. and Thompson, R. (1971). “Recovery of inter-block information when block sizes are unequal.” *Biometrika*. 58, 545 – 554.
- Pearson, K. (1901). “On lines and planes of closest fit to systems of points in space.” *Philosophical Magazine*. 2, 559 – 572.
- R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

- Reilly, C. (2006). *Selenium in Food and Health*. 2<sup>nd</sup> ed. New York, NY: Springer.
- Revelle, W. (2015). *psych: procedures for personality and psychological research*. Northwestern University, Evanston, Illinois, U.S.A. <http://CRAN.R-project.org/package=psych>. Version = 1.5.8.
- Richardson, E.A., Pearce, J., Mitchell, R. and Shortt, N.K. (2013). “A regional measure of neighbourhood multiple environmental deprivation: relationships with health and health inequalities.” *The Professional Geographer*. 65 (1), 153 – 170.
- Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Saisana, M., Saltelli, A. and Tarantola, S. (2005). “Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 168 (2), 307 – 323.
- Saltelli, A. (2000). *What is Sensitivity Analysis?* In: Saltelli, A., Chan, K. and Scott, E.M. (eds). *Sensitivity Analysis*. Chichester: Wiley.
- Saltelli, A. and Annoni, P. (2010). “How to avoid a perfunctory sensitivity analysis.” *Environmental Modelling & Software*. 25, 1508 – 1517.
- Saltelli, A., Chan, K. and Scott, E.M. (eds) (2000). *Sensitivity Analysis*. Chichester: Wiley.
- Scottish Environment Protection Agency (2014). <http://www.sepa.org.uk>.

Scottish Environment Protection Agency (2015).

<http://apps.sepa.org.uk/spria/Pages/SubstanceInformation.aspx?pid=1>.

Scottish Government (2009). *Scottish Soil Framework*.

<http://www.scotland.gov.uk/Publications/2009/05/20145602/0>. © Crown copyright.

Scottish Government (2012). *Scottish Index of Multiple Deprivation*.

<http://www.simd.scotland.gov.uk/publication-2012>. © Crown copyright.

Scottish Natural Heritage (2009). <http://www.snh.gov.uk/docs/B424902.pdf>.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Chichester: Wiley.

Silva, L.T. and Mendes, J.F.G. (2012). “City Noise–Air: An environmental quality index for cities.” *Sustainable Cities and Society*. 4, 1 – 11.

Spearman, C. (1904). “General intelligence: objectively determined and measured.” *American Journal of Psychology*. 15, 201 – 293.

Steltner, H., Staats, R., Timmer, J., Vogel, M., Guttman, H.M. and Virchow, J.C. (2002). “Diagnosis of sleep apnea by automatic analysis of nasal pressure and forced oscillation impedance.” *American Journal of Respiratory and Critical Care Medicine*. 165 (7), 940 – 945.

Sugiura, N. (1978). “Further analysis of the data by Akaike’s information criterion and the finite corrections.” *Communications in Statistics – Theory and Methods*. A7, 13 – 26.

Suter, G.W. (2001). “Applicability of indicator monitoring to ecological risk assessment.” *Ecological Indicators*. 1, 101 – 112.

United States Environmental Protection Agency (1992). *Lead poisoning and your children (800-B-92-0002)*. Office of Pollution Prevention and Toxics. Washington, D.C.

United States Environmental Protection Agency (2012a). *Macroinvertebrates and habitat*. <http://www.water.epa.gov/type/rsl/monitoring/vsm40.cfm>.

United States Environmental Protection Agency (2012b). *Phosphorus*. <http://www.water.epa.gov/type/rsl/monitoring/vsm56.cfm>.

University of Edinburgh (2015). <http://www.geos.ed.ac.uk/homes/s0198247/variograms.html>.

Vallero, D.A. (2014). *Fundamentals of Air Pollution*. 5<sup>th</sup> ed. Amsterdam: Academic Press.

Vierra, A.J. and Garrett, J.M. (2005). “Understanding interobserver agreement: the kappa statistic.” *Family Medicine*. 37 (5), 360 – 363.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall.

World Health Organisation (2001). *Arsenic and arsenic compounds*. Environmental Health Criteria 224. The International Programme on Chemical Safety (IPCS). <http://www.inchem.org/documents/ehc/ehc/ehc224.htm>.